Moral Responsibility

The ways of scepticism

Carlos J. Moya

Routledge
Taylor & Francis Group

Moral Responsibility

We are strongly inclined to believe in moral responsibility, that some human agents truly deserve moral praise or blame for some of their actions. However, recent philosophical discussion has put this natural belief in the reality of moral responsibility under suspicion. There are important reasons to think that moral responsibility is incompatible with both determinism and indeterminism, possibly rendering moral responsibility an impossibility.

This book lays out the major arguments for scepticism about moral responsibility and subjects them to sustained and penetrating critical analysis. *Moral Responsibility* lays out the intricate dialectic involved in these issues in a helpful and accessible way. The book goes on to suggest a way in which scepticism can be avoided, arguing that an excessive pre-eminence given to the will might lie at the root of scepticism of moral responsibility. Carlos Moya offers an alternative to scepticism, showing how a cognitive approach to moral responsibility which stresses the importance of belief would rescue our natural and centrally important faith in the reality of moral responsibility.

Carlos J.Moya lectures in philosophy at the University of Valencia, Spain.

Routledge Studies in Ethics and Moral Theory

1. The Contradictions of Modern Moral Philosophy
Ethics after Wittgenstein
Paul Johnston

2. Kant, Duty and Moral Worth
Philip Stratton-Lake

3. Justifying Emotions
Pride and Jealousy
Kristján Kristjánsson

4. Classical Utilitarianism from Hume to Mill
Frederick Rosen

5. The Self, the Soul and the Psychology of Good and Evil
Ilham Dilman

6. Moral Responsibility
The ways of scepticism
Carlos J.Moya

Moral Responsibility

The ways of scepticism

Carlos J.Moya

Routledge
Taylor & Francis Group
LONDON AND NEW YORK

First published 2006 by Routledge 2 Park Square, Milton Park, Abingdon, Oxon, OX14 4RN

Simultaneously published in the USA and Canada by Routledge 270 Madison Avenue, New York, NY 10016

Routledge is an imprint of the Taylor & Francis Group

This edition published in the Taylor & Francis e-Library, 2006.

"To purchase your own copy of this or any of Taylor & Francis or Routledge's collection of thousands of eBooks please go to http://www.ebookstore.tandf.co.uk/."

British Library Cataloguing in Publication Data A catalogue record for this book is available from the British Library

Library of Congress Cataloging in Publication Data A catalog record for this book has been requested

ISBN 0-203-96560-4 Master e-book ISBN

ISBN10: 0-415-37195-3 (Print Edition)
ISBN13: 978-0-415-37195-7 (Print Edition)

Taylor & Francis Group is the Academic Division of T&F Informa plc.

For my dear Milagro, Nuria and Ana And to the memory of J.L.Blasco

Contents

Acknowledgements

I have been interested in the subjects of moral responsibility and free will for many years. During this time I have benefited from exchanges and conversations with many people. I cannot mention them all, but let me give special thanks to my students, both undergraduates and postgraduates. I have learnt from them more than I have taught them. I should also mention my colleague and friend Tobies Grimaltos, with whom I have been giving a postgraduate course on freedom and related issues for several academic years. I owe to him much stimulus and help, though I have not been able to convince him that determinism is not a self-evident and necessary truth. Let me also thank other colleagues in the Department of Metaphysics for many ways in which they have helped me to develop my ideas about free will and moral responsibility.

Writing this book was possible owing to a grant awarded in 2002 by the Spanish Ministry of Education within a scheme called (in English translation) "Grants for stays of Spanish University teachers in foreign and Spanish research centres, including the program 'Salvador de Madariaga'". So I am much indebted to the Spanish educational authorities for giving me this excellent opportunity, as well as to the University of Valencia for granting the leave I applied for, which allowed me to spend all the academic year 2002–3 in the Department of Philosophy of the University of Sheffield. I must thank the teaching and administrative staff of the Department for their help and hospitality. Special thanks are due to Christopher Hookway, who was responsible for my research and wellbeing during my stay and helped me at several times with his comments and advice about some parts of the manuscript. I am also grateful to Stephen Laurence for inviting me to present some of the work I had been doing so far in a session of the excellent weekly Seminars of the Philosophy Department. This was a good opportunity for giving wider publicity to my thoughts on free will and moral responsibility and improving my work. Several members of the Department made very good and helpful comments and criticisms of my paper, which included part of what is now the second chapter of this book. I am also grateful to Jennifer Saul and Leif Wenar, who read some parts of the manuscript and gave me good advice and encouragement.

I have also presented aspects of this book in various seminars and conferences and I have highly benefited from the reactions, comments and criticisms of the respective audiences. I am grateful to Saul Smilansky and Jonathan Dancy, from whom I received help and good comments during conferences in Lund and Granada. Many thanks are also due to Juan Acero, who invited me to present the main lines of my research on moral responsibility in a seminar that took place recently in the Philosophy Department of the University of Granada; and many thanks to the members of the Department who attended the seminar and offered good and useful remarks. I am also grateful for a seminar I gave in Valencia to other members of Phronesis, an analytic philosophy group. Lino San Juan and Marta Moreno deserve special thanks for their comments during this seminar. I am very grateful to Eduardo Ortiz for reading the whole manuscript and making wise remarks on it. I should also thank Raimo Tuomela for inviting me to take part in a one-

day symposium on action, freedom and responsibility which took place recently in Helsinki, where I met Alfred Mele and Robert Audi, as well as several postgraduate students in the Department of Social and Moral Philosophy and benefited from their papers and their comments on mine.

At a different but no less important level I have to express my gratitude to the persons who made my almost ten months in Sheffield an agreeable and enjoyable experience. First of all I have to thank my wife, Milagro, and my little daughter, Ana, who spent in Sheffield a large part of those months. Without their love, company and encouragement, writing this book would have been a much harder task, and I wonder whether it would have been possible at all. The stay in Sheffield was a very good period for us three. The fact that we lived in a beautiful Edwardian house, in a quiet and nice area, contributed a lot to our happiness and wellbeing. And we all have to thank Jo Hookway for having found this house and for many other ways in which she cared about us. She and her husband, Chris, were invariably good hosts and friends, and we keep a debt of gratitude to them. We are also grateful to the Overseas Wives Wednesday Club of the University of Sheffield, a certainly admirable institution. Joining it made an invaluable contribution to Milagro's (and thereby to Ana's and my own) happiness and contentment. Let us express our gratitude to some members of the staff. These include Mrs Rosemary Boucher, who encouraged Milagro to join the club and made her access to some related services possible, as well as Mrs Marion Maitlis, a lively and lovely lady whose kindness and warmth we will always remember. A younger Dutch lady, Arnolda Beynon, was also very kind with us, and we still remember the wonderful house she and her husband possess in the Peak District, where we enjoyed their company and a nice meal. Let me also mention some other persons who made the stay in Sheffield warmer and nicer. They include Jenny Saul and her husband, Ray, who offered us their house for an excellent dinner as well as for a nice party, and Rob Hopkins, who invited me to his house and cooked a memorable turbot he had bought in Birmingham's fish market. After Milagro and Ana had left Sheffield, and I started to feel a bit lonely, it was very important for me to enjoy the company of some Spanish friends. Antonio Peidro, an old friend of mine, spent some days in my house. And I will always remember the wonderful moments spent with him and with my new (and much younger) friends Esa Diaz, Alfredo Muyo and Tamara Ojeda.

This book falls within the scope of the research project BFF2003–08335– C03–01, awarded by the Spanish Ministry of Education.

Most of subsection "Actual blockage cases", in Chapter 2, was previously published in *Critica,* vol. 35, 2003, pp. 109–20. I thank the editors of this journal for allowing me to use this material.

Let me finally acknowledge the permissions that have been granted for reproducing extracts from the following works:

- *The Significance of Free Will* by Robert Kane, copyright 1996 by Robert Kane. Used by permission of the author and Oxford University Press, Inc.
- *Freedom Within Reason* by Susan Wolf, copyright 1990 by Oxford University Press. Used by permission of Oxford University Press.
- *The Importance of What We Care About* by Harry G.Frankfurt, copyright 1988 Cambridge University Press. Reproduced with permission of the author and publisher.

• *Responsibility and Control: a theory of moral responsibility* by John Martin Fischer and Mark Ravizza, copyright 1998 Cambridge University Press. Reproduced with permission of the publisher.
• *Freedom and Belief* by Galen Strawson, copyright 1986 by Galen Strawson. Used by permission of Oxford University Press.
• *An Essay on Free Will* by Peter Van Inwagen, copyright Peter Van Inwagen 1983. Used by permission of Oxford University Press.
• "Source incompatibilism and alternative possibilities" by Derk Pereboom, in D.Widerker and M.McKenna (eds) *Moral Responsibility and Alternative Possibilities,* 2003, Ashgate Publishing Company. Used by permission of the publisher.

Introduction

Scepticism about moral responsibility (SMR)

The main concern of this book is scepticism about moral responsibility. By "moral responsibility" I understand that property of human agents by virtue of which they truly and objectively deserve moral praise or blame for some of their actions. We are naturally non-sceptical about this property. Even philosophical sceptics tend to praise or blame people spontaneously for some of their actions, though they may try to recoil from these spontaneous reactions after reminding themselves of their own reflectively acquired scepticism. The actions for which we hold human agents to be truly morally praise- or blameworthy are those that we judge to be morally right or wrong and that we assume were up to them. "Up to them" means roughly that these actions are ultimately attributable to the agents as their sources or authors and that, with respect to such actions, they had available alternatives: they could have acted in a different way, or could simply have avoided acting as they did. We assume that sometimes, indeed rather often, these conditions for moral responsibility, understood in the sense of objective praise- or blame worthiness, are actually met by human beings. If we come to think that, at some particular occasion, they are not, we naturally soften or even withdraw our judgement.

But consider that being able to satisfy these conditions in some particular occasions is what we understand by being a free agent. These conditions, then, are central to our notion of freedom, or of free will. We are also naturally convinced that many—perhaps most—human beings are free agents, or have a free will in that sense: that they can be authors or sources of some of their actions and that, in relation to those actions, they could have done otherwise. We may also say, then, that, on the assumption that a particular agent is a free agent, we hold her truly morally responsible for a particular action provided that we believe that, in acting that way, she exercised such ability (met those conditions) and so acted freely. In other words, we naturally assume that freedom is a necessary condition of moral responsibility. This is why we refuse to ascribe moral responsibility to some animals, or to small children: we think they lack a free will or, to use a medieval expression, a *liberum arbitrium*.

These convictions are a common starting point for all those who begin to think philosophically about these matters. Without these basic, natural intuitions, the philosophical problems of free will and moral responsibility would not exist. These problems arise, however, when we start reflecting on what would have to be the case in order for these natural intuitions to be true. It soon appears, on reflection, that being a free agent, and so one who may objectively deserve moral praise or blame for some of her actions, is trickier than it appears from our natural, spontaneous point of view. At the end of this reflection, some thinkers may come to the sceptical conclusion that having a free will, and so being a morally responsible agent, is just not possible.

If, as seems initially true, free will is a necessary condition of moral responsibility, scepticism about the former implies scepticism about the latter. In fact, scepticism about free will has grown significantly in recent times. To mention only a few examples, such books as *The Non-reality of Free Will* (Double 1991), *Free Will and Illusion* (Smilansky 2000) or *Living Without Free Will* (Pereboom 2001), whose titles are already expressive enough of their content, bear witness to this increasingly sceptical stance about free will. Not surprisingly, the authors of these books are sceptical about moral responsibility as well. There have, however, been some attempts to prevent scepticism about free will from spreading to moral responsibility, by rejecting the view that free will, understood as freedom to choose or act otherwise, is actually required for moral responsibility. We shall refer to these attempts later in this book, and argue that they are not successful. However, holding that freedom to do otherwise is not required for moral responsibility is not enough for that purpose, unless one is also prepared to accept that being the true origin or author of some of our actions, by having control over their springs, is not required for moral responsibility either. Moreover, there is room to argue (correctly, in our view) that these two conditions, alternative possibilities and authorship or control, as they might be called, are not independent of one another, so that lack of alternatives undermines the degree of origination and control that would be required for moral responsibility.

The main source of scepticism about moral responsibility is, then, scepticism about free will, or about the freedom-relevant conditions of moral responsibility. This is the route towards scepticism about moral responsibility that we shall be investigating in this book.

From a historical point of view, belief in free will was soon perceived to be in tension with the possibility of a world, of which human beings are a part, governed by fate, or necessary natural laws, or the decrees of God. Necessity, in any of these forms, was widely felt to be threatening to human freedom. So, for example, both ancient Epicureans and Stoics came to think that, if the atoms, the ultimate constituents of all things, obeyed ineluctable laws, human freedom would not be possible: both alternative possibilities and control would be ruled out, and with them moral responsibility, understood as true desert. This is the first clear statement of what is presently known as incompatibilism, the thesis that determinism and freedom cannot coexist. While Epicureans attempted to leave room for human freedom by holding that atoms sometimes suffered uncaused and unpredictable swerves, thereby adopting a libertarian position, the Stoics instead embraced the doctrine of unrestricted natural necessity and, consequently, denied that human freedom was possible. They were, in today's terminology, hard determinists. Strong echoes of stoicism can be heard in the work of Spinoza, a prominent hard determinist.

The assumption that determinism is not compatible with freedom is natural and was generally taken for granted until fairly modern times. Even nowadays, non-philosophers tend to accept it as almost self-evidently true. But Hobbes and Hume called it into question by first advancing compatibilism. Hume presented compatibilism, the thesis that there is no contradiction, no incompatibility between determinism and freedom, as the solution to the venerable problem of the relationship between them. However, far from being generally accepted as an end-point to the controversy, compatibilism quickly became a third contender in the discussion, along with the two traditional forms of

incompatibilism, namely libertarianism, which holds that human freedom exists and that therefore determinism is false, and hard determinism, which sustains the opposite thesis.

Though in a large number of versions and with many nuances, these three broad positions can still be said to roughly define the field of contemporary discussion about free will and moral responsibility, and each of them has important representatives. However, a new character has appeared on the scene, namely the true, across-the-board sceptic, who holds that free will and moral responsibility are certainly incompatible with determinism, *but also with* indeterminism. Though the hard determinist can also be said to be a sceptic about free will, in so far as she believes that determinism is true and that it precludes free will, if determinism were false after all, then free will might be a real property of (some) human beings. The true, across-the-board sceptic, however, closes this crack as well.

This radical form of scepticism is a late fruit of a significant difference between the traditional and the contemporary discussion, namely that, unlike what was generally the case after the outbreak of modern mathematical natural science and especially of classical Newtonian physics, a strictly deterministic view of nature has ceased to be widely taken for granted, and the possibility that some basic physical processes may be indeterministic has been taken seriously. Some libertarian incompatibilists viewed, and still view, the probabilistic laws of quantum physics as the natural enabling condition of freedom that they were hoping for and as a support for their philosophical position. However, the suspicion quickly arose that, as early compatibilists had already suggested, indeterminism might be threatening to free will and moral responsibility. If, according to the deterministic picture, human choices and actions are inevitable outcomes of the past and the natural laws, then, according to incompatibilists, free will is undermined, for there seems to remain no room for either of its aspects, namely alternatives and deep origination or control. Since these, in turn, are the freedom-relevant necessary conditions for moral responsibility, this property loses its footing as well. But if human choices and actions are instead the result of unpredictable, random events in our brain at the subatomic level, then, even if alternative possibilities are possible, control over our choices between them, and the associated idea of our being true authors and sources of our actions, are no less effectively eroded, and with them free will and moral responsibility as well.

So the rise of a rigorous, natural-scientific, indeterministic view of the natural world has reshaped the contours of the philosophical problems of free will and moral responsibility. In one sense, it has worsened those problems rather than solving them. The traditional question of the compatibility between free will and a deterministic natural world has been enlarged so as to encompass the compatibility between free will and an *indeterministic* natural world as well. And a negative answer to both questions has given rise to the radical, across-the-board form of scepticism that we have referred to. The threat to the possibility of free will and moral responsibility does not come just from a *deterministic* natural world, but from the *natural world* as such, whether deterministic or not. And if those properties are shown to be incompatible with the natural world, whatever its ultimate structure may be, the suspicion arises that the concept of such properties is simply incoherent and so unable to be instantiated at all.

This radical, across-the-board form of scepticism about moral responsibility, on the basis of scepticism about its freedom-relevant conditions, will be the central theme of this essay. To proceed in an orderly, systematic way, we shall conceive of this form of scepticism as the conclusion of a very general sceptical argument that we shall dub "SMR" (scepticism about moral responsibility). This argument is an abstract, simplified reconstruction out of several positions held in contemporary debates about moral responsibility and free will. But some closely related arguments can also be found in an explicit form in some recent works. It may be useful to look at some of them before formulating SMR. Common to all these arguments, including SMR, is a disjunctive premise asserting that either determinism holds or it does not.

An argument of this sort can be found explicitly formulated in a recent paper by Peter Unger. It concerns free will rather than moral responsibility, and gives expression to what nowadays, in Unger's words, "may be the real heart of 'the problem of free will'" (Unger 2002:4). The argument is as follows:

First Premise: If Determinism holds, then, as everything we do is inevitable from long before we existed, nothing we do is anything we choose *from available alternatives* for our activity.

Second Premise: If Determinism *doesn't* hold, then [while some things we do may be inevitable from long before our existence and, as such, it's never within our power to choose them for ourselves] it may be that some aren't inevitable—but, as regards any of these others, it will be a *matter of chance* whether we do them or not, and, as nothing of *that* sort is something we *choose* to do—nothing we do is anything we choose from available alternatives for our activity.

Third Premise: Either Determinism holds or it doesn't.

Therefore,

Conclusion: Nothing we do is anything we choose from available alternatives for our activity.

(Unger 2002:4)

This is a sceptical argument about free will understood in terms of choice between alternative possibilities, as freedom to do otherwise. It is silent, however, about free will understood in terms of control and origination. As a result, it is not clear why, from the fact that whether we do something or not is a matter of chance, we should infer that we do not *choose* (in the relevant sense) at all. Moreover, as it stands, this argument does not threaten moral responsibility. For it to do so, an additional premise would be needed, to the effect that choosing from available alternatives is a requirement for moral responsibility. Unger's argument, however, manifests a consciousness of the new shape that the old problem of free will has taken in recent times, and of the radical form of scepticism it has given rise to. We shall see how this new shape and this radical sceptical stance affect the question of moral responsibility as well. This question has a wider scope than the question of free will, since it encompasses the latter as well. In this broader context, some of Unger's contentions in his sceptical argument will receive further illumination and support.

Van Inwagen (2000) also considers a similar sceptical argument about free will. He thinks that compatibilism, the thesis that free will is compatible with determinism, is implausible, but he adds that free will "also seems to be incompatible with indeterminism". Though he is a libertarian, not a sceptic, and thinks that "free will undeniably exists", he sees the strength of the scepticism that is thereby generated, and his conclusion in that paper reflects his puzzlement: "I conclude that free will remains a mystery—that is, that free will undeniably exists and that there is a strong and unanswered *prima facie* case for its impossibility" (Van Inwagen 2000:1–2).

On the basis of the preceding considerations, let us now proceed to formulate our own sceptical argument, SMR. In the simple, canonical form in which we propose to construe and deal with it in this book, the argument runs as follows:

SMR (Scepticism about moral responsibility):

A. Either determinism is true or it is not true.
B. If determinism is true, moral responsibility is not possible.
C. If determinism is not true, moral responsibility is not possible.
D. Therefore moral responsibility is not possible.

SMR is patently valid. Whether it is sound, and so whether it establishes the truth of its sceptical conclusion, will thus depend upon the truth of its premises. Premise B is the traditional incompatibilist thesis as applied to determinism and moral responsibility rather than free will. Premise C expresses the view that moral responsibility is also incompatible with indeterminism. The argument thus reflects the radical, across-the-board form of scepticism we have been talking about, as the sign of—paraphrasing Unger—the real heart of the problem of moral responsibility in present times.

The structure of this book is closely related to the structure of SMR itself. Given that SMR is logically valid, the book is concerned, in its first four chapters, with the reasons for thinking that its premises are true. After commenting rather briefly on premise A, the first chapter embarks on the discussion of premise B. Chapters 2 and 3 are devoted to a further discussion of this premise, while Chapter 4 deals with premise C. This task involves a rather long perambulation through large areas of contemporary debates about moral responsibility and free will. The conclusion of these four chapters is that the reasons for the truth of SMR's premises are very powerful and that, consequently, the case for SMR's sceptical conclusion about moral responsibility is also very strong. The fifth, and final, chapter is an attempt to resist this conclusion by showing a way in which one of SMR's premises, namely premise C, might be questioned.

More precisely, a tree trunk and its roots could represent the structure of this book in its first four chapters. Thus the book has a ramified structure. Following this metaphor, the trunk corresponds to SMR's sceptical conclusion, namely that moral responsibility is not possible. The trunk is supported by three thick roots, the three premises of SMR, each of which is necessary, and all of them jointly sufficient, for the trunk to stand firmly in place. While premise A is mainly self-supporting, premises B and C need additional support. Each of these two premises is the conclusion of further arguments and is supported by their premises.

Premise B is the conclusion of two independent arguments, each of which, if sound, is sufficient for its truth. The first of these two arguments, which we dub "the Incompatibilist Argument", has the following two premises: 1) Determinism rules out alternative possibilities of decision and action; and 2) alternative possibilities are necessary for moral responsibility. The premises of the second argument are: 1) Determinism rules out ultimate control over our actions; and 2) ultimate control is necessary for moral responsibility. The conclusion of either argument is SMR's premise B, namely that, if determinism is true, moral responsibility is not possible, or, in other words, that determinism is incompatible with moral responsibility.

The first premise of the Incompatibilist Argument is discussed in Chapter 1, viewing it as the conclusion of several arguments, the main one of which is the so-called Consequence Argument. The second premise of the Incompatibilist Argument is in turn dealt with in Chapter 2. We argue for the truth of this premise mainly in a negative way, trying to show that none of the attacks on it, of which Frankfurt's has been the most influential, is finally successful.

The premises of the second argument for SMR's premise B are discussed in the third chapter. The first of these premises, namely that determinism is incompatible with ultimate control, is largely taken for granted, on the basis that it follows immediately from the very concept of determinism. More contentious is premise 2, according to which ultimate control is required for moral responsibility. This premise is the main subject of Chapter 3. We argue for its truth on the basis that the main approaches to moral responsibility that attempt to dispense with ultimate control can thereby be shown to be ultimately flawed.

The outcome of these three chapters is that SMR's premise B has strong evidence in its favour.

SMR's premise C is the object of the fourth chapter. The premise asserts the incompatibility of indeterminism with moral responsibility. It will be seen, again, as the conclusion of a further argument, and so as rooted in, and supported by, its premises. This argument has one premise in common with the second argument for SMR's premise B. This common premise states that ultimate control over our decisions and actions is necessary for moral responsibility. Since this premise has already been discussed, and accepted, in Chapter 3, it is now largely taken for granted. The second premise asserts that indeterminism rules out control, and *a fortiori* ultimate control, over our decisions and actions. This premise is itself the conclusion of the so-called "Mind" argument, which is dealt with in several versions of it. The result of this chapter is that SMR's premise C has very strong support as well.

The conclusion of these four chapters is, then, that SMR is a very powerful sceptical argument and that the possibility that its conclusion, the impossibility of moral responsibility, is true should be taken very seriously.

In the fifth, and final, chapter we explore a way in which SMR's sceptical conclusion could be resisted. The ramified dialectical structure depicted in the preceding chapters allows for a quite perspicuous overview of the logical relations of dependence between the elements that support scepticism about moral responsibility and of the ways that lead to it. The elements of this structure are very tightly put together. This tightness gives the structure its strength, but it is also a source of its potential weakness, for the failure of even a slender root might be sufficient for the sceptical trunk to fall. As we have pointed

out, SMR's premise B is the conclusion of two independent arguments that support it. One of them starts from the necessity of alternative possibilities for moral responsibility and the other from the necessity of ultimate control. Together with the contention that these conditions are incompatible with determinism, each argument leads to premise B as its conclusion. However, only one argument that starts from the necessity of ultimate control for moral responsibility leads to SMR's premise C through the contention that ultimate control is incompatible with indeterminism. So one result of a general inspection of the dialectical structure is that, at least formally, premise C is SMR's weakest link (given that premise A is logically necessary). Another interesting result is that the argument that leads to premise C as its conclusion has one premise in common with the second argument for premise B, namely that ultimate control is necessary for moral responsibility.

Putting these two results together suggests, first, that the necessity of ultimate control for moral responsibility plays a central role in supporting the whole dialectical structure, and, second, that rejecting this condition would directly undermine SMR's premise C, which in turn would undermine SMR's sceptical conclusion. This has, in fact, been a compatibilist move to defend the possibility of moral responsibility. This move is even more tempting for incompatibilists, since, unlike compatibilists, they do not need to reject, or reinterpret, the alternative possibilities condition in order to avoid scepticism. Finally, rejecting ultimate control becomes even more tempting given that some thinkers (notably Galen Strawson) argue that this condition makes an impossible demand, which, together with its necessity for moral responsibility, leads directly to scepticism about the latter.

However, we do not recommend this route. Its rapid anti-sceptical results would be bought at too high a price, for there are strong reasons to think that moral responsibility, understood as true, objective desert, as true praise-or blameworthiness, would not survive the rejection of some form of deep, ultimate control over our actions.

We said that SMR's premise C was the conclusion of two premises, the first being the necessity of ultimate control and the second the incompatibility of this condition with indeterminism. Our proposal, in essence, is to undermine premise C, thereby undermining SMR's sceptical conclusion, on the basis of rejecting the second of those two premises instead of the first. So we shall try to show that deep, ultimate control over our actions, though incompatible with determinism, can none the less be compatible with indeterminism.

The main consideration against this compatibility, which is the core of the "Mind" argument, is that indeterminism turns our decisions and actions into arbitrary, chancy events, so depriving us of control, and especially of rational control, over them. We argue, however, that this consideration is powerful, and perhaps decisive, if ultimate control, and moral responsibility itself, is taken to rest centrally on will-related acts, especially choices. It is also this conative approach, as it might be called, that makes ultimate control appear to make an impossible demand. In its place we recommend a cognitive approach to moral responsibility and its freedom-relevant conditions, namely alternative possibilities and ultimate control. According to the recommended approach, the central aspect of free will, and of moral responsibility, is not choice but belief. A particular group of beliefs, with an evaluative content, is especially relevant. We argue that we can have a sort of control over our beliefs, including our evaluative beliefs, that is

not based on choice and that is none the less deep enough to satisfy the intuitions that underlie the condition of ultimate control. We also contend, on the basis of this cognitive approach, that indeterminism need not deprive us of rational control over our actions and their cognitive springs. To this end, we distinguish two perspectives on the place and role of indeterminism in practical rationality and argue that, though one of them ("bottom-up indeterminism", as we call it) may be damaging for rational control, the other ("top-down") need not be so. On the contrary, it may be constitutive of that control. On the basis of our recommended cognitive approach and of a "top-down" view of indeterminism, we contend that the "Mind" objection to libertarian incompatibilism can be successfully answered, thus clearing the way for a rejection of SMR's premise C and its sceptical conclusion.

Though this book aims mainly at a theoretical understanding of its central topics, it is not intended to be without practical consequences. It is our hope that it may help us, on the basis of that understanding, to develop and enrich our freedom and the quality of our life. However, these practical consequences will remain largely implicit.

This book allows for different uses. As a whole, it is a monographic essay about the subject referred to by its title. However, given its internal structure, it can also be used as an advanced textbook about moral responsibility and free will, and provide an overview of this wide and rather intricate field. Moreover, each of its chapters, with the exception of the fifth, which presupposes knowledge of the rest, can be read and used separately for courses or seminars about the subjects indicated by their titles.

1
Determinism and alternative possibilities
(SMR's premises A and B)

Remember the sceptical argument about moral responsibility:

SMR (Scepticism about moral responsibility):
A. Either determinism is true or it is not true.
B. If determinism is true, moral responsibility is not possible.
C. If determinism is not true, moral responsibility is not possible.
D. Therefore moral responsibility is not possible.

In this chapter, we shall briefly examine what appears to be the least contentious premise of SMR, namely premise A. Afterwards, we shall start evaluating premise B, which asserts the incompatibility between determinism and moral responsibility (the thesis known as "incompatibilism"). An important argument for the truth of this premise may be called the Incompatibilist Argument. It runs as follows:

1) Moral responsibility requires alternative possibilities: an agent is morally responsible for an action of hers only if she could have done otherwise. 2) Determinism rules out alternative possibilities: if determinism is true, nobody could have done otherwise than she in fact did. 3) Therefore, if determinism is true, moral responsibility is not possible.

The conclusion of the Incompatibilist Argument is SMR's premise B. So the premises of the Incompatibilist Argument are directly relevant to the truth of SMR's premise B. In this chapter, we shall comment on premise 2. This chapter will have to include some formal arguments, which we shall try to keep to a minimum. The rest of the book will dispense with formal arguments and proceed in terms of natural language.

Determinism (SMR's premise A)

Apparently, SMR's premise A is not problematic. It looks like an instance of the general scheme "*p* or not *p*", and instances of this scheme are logically necessary truths. But some considerations are in order. In the scheme "*p* or not *p*", the variable *"p"* is supposed to range over propositions or sentences with a definite and truth-evaluable content or meaning. If a presumptive instance of this scheme does not satisfy this condition, one need not accept its truth. The question, then, is whether the thesis referred to by "determinism" has a definite and truth-evaluable content. Not everybody accepts this. In a famous article, Peter Strawson said he belonged to "the party of those who do not know what the thesis of determinism is", but he went on to admit that "though darkling, one has

some inkling—some notion of what sort of thing is being talked about" (Strawson 1962:59). Even in the light of the current definitions of "determinism" that can be found in the literature on free will and moral responsibility, Strawson's reticence about the content of that thesis is understandable, for those definitions make explicit or implicit use of some doubtful notions. According to Van Inwagen, for example, determinism is…

> [T]he conjunction of these two theses:
>
> For every instant of time, there is a proposition that expresses the state of the world at that instant;
>
> If *p* and *q* are any propositions that express the state of the world at some instants, then the conjunction of *p* with the laws of nature entails *q*.
>
> This definition seems to me to capture at least one thesis that could properly be called "determinism". Determinism is, intuitively, the thesis that, given the past and the laws of nature, there is only one possible future. And this definition certainly has that consequence.
>
> (Van Inwagen 1983:65)

According to determinism, so defined, the conjunction of the proposition that expresses the state of the world at a certain instant and the proposition that expresses the laws of nature logically entails a proposition that expresses the state of the world at any other instant. This entailment goes from the past to the future and vice versa, though it is the first entailment that usually comes naturally to one's mind and is usually emphasized in the literature. Van Inwagen himself insists on the past-to-future relation in a shorter definition of "determinism" in the same work: "*Determinism*…is the thesis that there is at any instant exactly one physically possible future" (Van Inwagen 1983:3). Though laws of nature are not mentioned in this definition, they are implicitly introduced when he says that, given the state of the world at a certain instant, only one future is "physically possible". Though many futures are logically possible given the state of the world at a certain instant, only one of them is physically possible, or, in other words, only one of them is logically possible *given* also the natural laws. In a recent article, Ted Warfield defines "determinism" as follows: "Determinism is the thesis that the conjunction of the past and laws implies all truths" (Warfield 2000:173). And Ekstrom conceives it as the thesis that "at any particular moment, there is, given the actual past and the laws of nature, exactly one way the world could go" (Ekstrom 2000:16), a definition she takes to be equivalent to Van Inwagen's.

This sample of current conceptions of determinism is enough to give a sense of the difficulties involved in the contention that this thesis has a definite and truth-evaluable content. Take Warfield's definition, for example. Of course, the *past* is not the sort of thing that can imply a truth. Only a proposition that *describes* the past can do that. And if this proposition, together with the laws of nature, is to imply *all* truths, it has to be a *complete* description of the past. Warfield, then, makes implicit use of the notion of a *complete description* of the past. This notion also features implicitly in Ekstrom's definition, and Van Inwagen explicitly appeals to an equivalent idea when he talks about a "proposition that expresses the state of the world" at a certain instant. However, it is not clear what a *complete description* of the past or of the state of the world (at a certain instant) can be. And if this notion lacks a reasonably definite content, this will infect the

thesis of determinism as well, thus compromising the truth of SMR's premise A and preventing the argument from getting off the ground.

Furthermore, if the conjunction of a complete description of the past and the natural laws is to imply any proposition, the description has to be made in the vocabulary in which the laws are stated. Moreover, the laws should not allow for any exceptions: they should be strictly deterministic, not probabilistic, and such laws are likely to be found only in basic physics. In view of all this, it seems reasonable to require that the description of the past be made in the vocabulary of physics, that it be a *physical* description. Some sense can then be made of the idea of a complete description if it is understood as a complete physical description, say as a description of the positions and velocities of elementary particles in the universe at a certain instant in the past. As for the laws, they have to be non-probabilistic, but at least some physical laws are widely thought to be probabilistic. Let us assume, however, for the sake of argument, that they are not. On these assumptions, it is at least conceivable that a complete physical description of the state of the universe at an instant in the past, together with deterministic physical laws, implies all truths about the *physical* state of the universe at a later instant.

However, the sorts of propositions to be derived from that conjunction which are relevant to questions about moral responsibility are propositions about mental and intentional states and events that people can be in or bring about, such as desires, beliefs, choices or intentions, as well as intentional actions. So, in order for these propositions to be derivable from the conjunction of a complete physical description of the past and the natural laws, we need reliable nomological connections between physical and mental concepts or properties: we need something like psychophysical type-identity, or at least strong supervenience of mental properties on physical ones.

At least some of the above suppositions are highly problematic. But, for what concerns its relation to moral responsibility, determinism could survive the falsity of at least some of them. It might well be, for example, that the physical laws that hold at a subatomic level are probabilistic, but that indeterminacies at this level would be cancelled at higher ("macrophysical") levels of the organization of matter, such as atomic or molecular levels, so that atoms or molecules actually behaved according to strictly deterministic laws. If they did, it might still be that a complete macrophysical description of the state of the universe at a certain instant, together with macrophysical deterministic laws, would imply all macrophysical truths about the state of the universe at a later instant. This would still be a thesis recognizable as determinism. And, were it the case that mental properties supervened on macrophysical properties, determinism would then be true from the macrophysical level onwards.

Suppose, however, that microphysical indeterminacies are reflected or amplified, rather than cancelled, at the macrophysical level. Assuming the supervenience of all non-physical properties on physical ones, determinism would then be false at all levels, including the psychological level. It might also be false at the psychological level if mental properties did not actually supervene on (micro- or macro-) physical properties. But think that we do not need the *truth* of determinism in order for premise A of SMR to be true. Nor do we need determinism to be verifiable (it most probably is not). As we said earlier, all that is required is that the thesis of determinism should have a definite and truth-evaluable content. Now, in the light of the preceding discussion, the thesis of

determinism clearly seems to have the required content. And if it does, then premise A of SMR is true, and necessarily so.

SMR's premise B is much more contentious. It asserts incompatibility between determinism and moral responsibility. Let us start examining the reasons for thinking that this premise is true. One important reason is the Incompatibilist Argument, the conclusion of which is precisely the incompatibility between determinism and moral responsibility. According to premise 2 of the Incompatibilist Argument, determinism precludes alternative possibilities. Let us examine this contention.

Does determinism preclude alternative possibilities?

At first sight, the answer to this question would seem to be obviously affirmative. One would tend to agree immediately with Gary Watson when he writes: "If determinism is true, then clearly, in some sense, there are no alternative possibilities. Relative to the laws of nature and antecedent conditions, it is not possible that one does anything but what one does" (Watson 1987:154). However, proving this apparently obvious thesis has shown itself to be a rather complicated matter. A central argument for the truth of this thesis is known in the literature as the Consequence Argument. Its main proponent, Peter Van Inwagen, presents this argument informally as follows:

> If determinism is true, then our acts are the consequences of the laws of nature and events in the remote past. But it is not up to us what went on before we were born, and neither is it up to us what the laws of nature are. Therefore, the consequences of these things (including our present acts) are not up to us.
>
> (Van Inwagen 1983:56)

This certainly looks like a powerful argument, but the discussion about it is still going on. The main point of discussion is the assumption that, if an agent has no choice about the truth of a proposition *p,* and has no choice about the fact that *p* implies *q,* she also has no choice about *q*. This assumption is a Principle of Transfer of Powerlessness: it allows us to go from our lack of choice about (or power over) the past and the natural laws to our lack of choice about our present actions, via our lack of choice about the fact that, given determinism, these actions are entailed by the conjunction of a proposition that describes the state of the world at an instant in the (remote) past and the laws of nature. The conclusion of the argument is a conditional statement according to which, if determinism is true, we have no choice about which action we perform, or, in other words, if determinism is true, we cannot act otherwise than the way we actually do. Determinism, then, precludes alternative possibilities, which is premise 2 of the Incompatibilist Argument.

Van Inwagen presents three formal expositions of the argument, which he claims are equivalent, so that they stand or fall together. The Principle of Transfer of Powerlessness features in all three, either as an implicit assumption of some of the premises or as a separate rule of inference. For purposes of discussion, we shall consider the first and third formal expositions.

Let us start with the first formal exposition, or, in Van Inwagen's terms, the First Formal Argument. Suppose there is an agent, call her "J", who, at a certain instant, T, did not raise her hand. Let "T_o" denote a certain instant of time before J's birth, "P_o" denote a proposition that expresses the state of the world at T_o, "P" denote a proposition that expresses the state of the world at T, and "L" denote the conjunction into one proposition of all natural laws. Finally, let us say that an agent can render a certain proposition false just in case she can act so as to ensure the falsity of that proposition. Let us quote Van Inwagen:

The First Formal Argument consists of seven propositions, the seventh of which follows from the first six: (1) If determinism is true, then the conjunction of P_o and L entails P. (2) It is not possible that J has raised his hand at T and P be true. (3) If (2) is true, then if J could have raised his hand at T, J could have rendered P false. (4) If J could have rendered P false, and if the conjunction of P_o and L entails P, then J could have rendered the conjunction of P_o and L false. (5) If J could have rendered the conjunction of P_o and L false, then J could have rendered L false. (6) J could not have rendered L false. (7) If determinism is true, J could not have raised his hand at T.

(Van Inwagen 1983:70)

Obviously, there is nothing special in the action of raising one's hand, so the argument generalizes to any action that any agent performs at any time. Van Inwagen points out that the conditionals in 1–7 are material conditionals. Note that, in this argument, premise 4 presupposes the truth of the Principle of Transfer of Powerlessness. Remember that, according to this principle, if an agent has no choice about the truth of *p* and no choice about the fact that *p* implies *q,* she has no choice about the truth of *q* either. Premise 4 applies the contraposition of this principle, according to which, if an agent has a choice about the truth of *q,* and *q* is implied by *p,* she has a choice about the truth of *p* as well.

The Third Formal Argument also employs "P_o" and "L", though this time they abbreviate sentences that express the propositions that those symbols denoted in the First Argument. "P" abbreviates a sentence that expresses any true proposition. The symbol "$\Box$" represents broad logical necessity, and "$\rightarrow$" the material conditional. The argument makes use of a sentential operator, N. If P is a sentence, "N P" is to be read as "P, and no one has, or ever had, any choice about whether P", where someone's having a choice about a true proposition, P, is to be understood as her being able to act so as to ensure the falsity of P. Finally, the argument makes use of two inference rules, Alpha and Beta. According to rule Alpha, $\Box$P implies N P. According to Beta, N P and N (P$\rightarrow$Q) implies N Q.

On this third presentation, the Consequence Argument runs thus (cf. Van Inwagen 1983:94–5):

1 . $\Box((P_o \& L) \rightarrow P)$	Consequence of Determinism
2. $\Box(P_o \rightarrow (L \rightarrow P))$	1, by sentential and modal logic
3. $N(P_o \rightarrow (L \rightarrow P))$	2, Alpha
4. $N\ P_o$	Premise (fixity of the past)
5. $N(L \rightarrow P)$	3, 4, Beta
6. N L	Premise (fixity of the laws)
7. N P	5, 6, Beta

So, if the Consequence Argument is sound, then, given determinism, no one has any choice with respect to any true proposition, including propositions about actions performed by human beings. Determinism is incompatible with alternative possibilities of action, with freedom to do otherwise. Note that, on this presentation, the Principle of Transfer of Powerlessness appears in an explicit form as the inference rule Beta.

As Van Inwagen himself acknowledges (cf. Van Inwagen 1983:96), rule Beta is the more contentious link in the argument. And criticisms of the Consequence Argument have usually attacked this link, in a more or less direct way. But there have been attempts to reject some premises of the argument as well. Let us now look at some of these criticisms.

Alternative possibilities: conditional analyses

The tension between determinism and alternative possibilities was widely felt before the Consequence Argument appeared. It seemed to many thinkers that, if determinism were true, nobody could have done otherwise than she in fact did. This conditional statement is, in fact, equivalent to the conclusion of the Consequence Argument. In a compatibilist vein, some thinkers tried to show that this statement was false, so that, contrary to appearances, determinism is compatible with alternative possibilities. According to the most influential proposal in this direction, which, as J.L. Austin reports (cf. Austin 1970), we owe to George Moore (though, in fact, it can already be found, in less elaborated forms, in Hobbes or Hume), a proper analysis of such statements as "S could have done otherwise" shows that they can be true even if determinism holds, so that there is no real incompatibility between alternative possibilities and determinism. According to this proposal, the claim that S could have done otherwise can be correctly analysed as the claim that S would have done otherwise if she had chosen (decided, tried) to do so. The truth of this conditional is compatible with the truth of determinism, as will be the truth of "S could have done otherwise" if the analysis is actually correct.

Suppose that this conditional analysis is correct, so that the proposition expressed by "S could have done otherwise" is actually equivalent to that expressed by "If S had chosen (decided, tried) to do otherwise, S would have done otherwise." On this supposition, the Consequence Argument fails to establish its conclusion, as compatibilists

have rightly insisted. One way of seeing this (cf. Kane 1996:47) is to focus on premise 4 of the First Formal Argument, which, as we pointed out, is a particular case of a Principle of Transfer of Powerlessness. The premise, as we saw, is as follows: "If J could have rendered P false, and if the conjunction of P_o and L entails P, then J could have rendered the conjunction of P_o and L false." One way in which J could have rendered P false is by raising her hand at T. So let us replace P by the proposition expressed by "J does not raise her hand (at T)." The first conjunct in the antecedent of premise 4 would now read "J could have rendered the proposition expressed by 'J does not raise her hand (at T)' false." Putting this in more colloquial terms, we simply have "J could have raised her hand (at T)."

Let us now express this last sentence in terms of the proposed conditional analysis. We have, then: "If J had chosen (decided, tried) to raise his hand (at T), she would have done so." Now, this is most likely true, even if determinism holds. And, provided that determinism holds, the second conjunct in the antecedent, namely that the conjunction of P_o and L entails P, will also be true. On the conditional analysis, then, the antecedent of premise four in the First Formal Argument will be true.

The consequent of this premise is: "J could have rendered the conjunction of P_o and L false." In terms of the proposed analysis, this is equivalent to: "If J had chosen (decided, tried) to render the conjunction of P_o and L false, she would have done so." Now, whether or not determinism holds, this clearly seems to be false. Even if J had chosen, decided or tried to modify either the past or the natural laws, or both, she would have failed to do so. On the conditional analysis, then, the consequent of premise 4 is false.

But, as Van Inwagen insists, the conditionals in the First Formal Argument are material conditionals. So, on the proposed analysis, the conditional that constitutes premise 4 has a true antecedent and a false consequent. The premise is false, therefore, as well as the Transfer Principle on which it is based, and the incompatibilist conclusion of the argument might be false as well.

The conditional analysis of "could have done otherwise" provides thus a way of resisting the conclusion of the Consequence Argument. Incompatibilists might reject the conditional analysis by arguing that premise 4 and the Transfer Principle are clearly more plausible than the analysis itself, but compatibilists would then accuse them of begging the question. Incompatibilists, of course, could also raise this accusation against their opponents, and the result would be a dialectical impasse or stalemate.

However, the stalemate might finally be broken in favour of the incompatibilists, for, independently of a prior commitment to the truth of premise 4 or of the Transfer Principle, there seem to be quite strong reasons for thinking that the conditional analysis is not correct after all. Famously, about forty years ago, Roderick Chisholm raised an important criticism of this analysis (cf. Chisholm 1964). Let A be the proposition expressed by "S could have done otherwise" and let B be the proposition expressed by "If S had chosen (decided, tried) to do otherwise, S would have done otherwise." According to the conditional analysis, A and B are logically equivalent. Now, if A and B were logically equivalent, A could be deduced from B (and vice versa). But A cannot be deduced from B. It might well be that B is true while A is false, provided that S could not have chosen to do otherwise. So A cannot be deduced from B alone, but from B plus the proposition expressed by "S could have chosen to do otherwise" (C). A, then, is not

equivalent to B, but, if at all, to B plus C. However, any attempt to analyse C itself in conditional terms will run into another proposition, D (say, "S could have chosen (decided, tried) to choose to do otherwise"), and an infinite regress will have started.

Donald Davidson advanced a proposal to rescue the conditional analysis (cf. Davidson 1973). According to him, the difficulty raised by Chisholm's criticism applies to any attempt at analysing "S could have done otherwise" in terms of a conditional whose antecedent contains a verb suggesting the idea of an *activity* on the agent's part. This happens with "choose", but also with "try", "decide", "intend" and similar verbs. The difficulty might then be sidestepped by including in the antecedent of the conditional a verb that does not suggest any activity. As Davidson writes: "The only hope for the causal analysis is to find states or events which are causal conditions of intentional actions, but which are not themselves actions or events about which the question whether the agent can perform them can intelligibly be raised" (Davidson 1973:147). Davidson's suggestion is that the most eligible states or events would be reasons, which, in Davidson's view, are pairs of beliefs and desires that causally and rationally explain an agent's actions. Beliefs and desires satisfy Davidson's requirement that the question whether agents can *perform* them is not intelligible.

As I understand it, Davidson's proposal would amount to the following. Instead of analysing A in terms of B plus C, which leads to an infinite regress of acts that the agent should be able to perform in order for her to be able to do otherwise, let us analyse A in terms of B*: "If S had had sufficient reasons to do otherwise, S would have done otherwise," where these reasons are appropriate beliefs and desires. Now, it might still be the case that B* were true while A was false, provided that the agent could not have had such sufficient reasons to do otherwise, so that A is not equivalent to B*, but, if at all, to B* and C*: "S could have had sufficient reasons to do otherwise." However, C*, unlike C, does not say that the agent could have performed a certain act (such as choosing, or deciding, or trying, etc.) and so it does not lead to an infinite regress. The conditional analysis may therefore stop with C* or, alternatively, it can go on with causal conditions that should have been met in order for the agent to be able to have had those reasons. Given determinism, this causal chain may well be potentially infinite, but this regress is not vicious, in that it does not require that the agent be able to perform an infinite number of acts. It is just the sort of sequence that begins to form when we ask for the causes of any event and that we usually bring to a halt by finding an event that, depending on the particular context, we find satisfactory as an explanation.

Is this version of the conditional analysis of "S could have done otherwise" correct? For it to be so, A should be equivalent to B* plus C*, which means that A could be derived from B* and C* (and vice versa). Contrary to appearances, however, this derivation is not correct, in that it contains a fallacy of equivocation concerning the term "could". In C*, assuming determinism (as one should, since the analysis is intended to show the compatibility of alternative possibilities and determinism) "could" means roughly "it is logically possible that". But, in A, "could" means, also roughly, "it was in her power" or "she was actually capable of". Bear in mind that we are dealing with the problem of moral responsibility. And in this context, when we ascribe moral responsibility to an agent on the basis that she could have done otherwise, we are not saying merely that her performing an alternative action was logically possible, but that she was actually capable of performing it. It is logically possible that I fly without

mechanical aid, but I am not actually capable of doing so. It would be unreasonable to ascribe moral responsibility to me for something I did on the basis that I could have flown without mechanical aid, even if it was logically possible. In A, then, "could" has a different (and stronger) content than in C*, which means that A cannot validly be derived from C* (plus B*). The equivocation fallacy might be avoided by giving to the term "could", in C*, the same meaning as it has in A. But then C* becomes false: in the sense in which I am actually capable of, say, drinking a glass of water, I am not actually capable of having different beliefs and desires from the ones I now have. Beliefs and desires are not the sort of states that are under our direct voluntary control.

The result of this discussion is that conditional analyses of "could have done otherwise" fall far short of undermining the Consequence Argument.[1] But the argument has been criticised in other ways. Of these, direct criticisms of rule Beta are probably the most powerful. Let us move on to them.

The Consequence Argument and rule Beta

Let us focus on the third formal presentation of the Consequence Argument, or the Third Formal Argument, in Van Inwagen's terms. An essential component of this argument is rule Beta, which licenses the inference from N P and N(P→Q) to N Q. As we can see, rule Beta allows us to conclude, starting from our lack of choice about a true proposition, our lack of choice about a logical consequence of that proposition, via our lack of choice about this relation of logical entailment. It is, then, a Principle of Transfer of Powerlessness. Remember that having no choice about a true proposition is understood as being unable to act so as to ensure the falsity of that proposition. Van Inwagen acknowledges that this is the weakest link in the Consequence Argument. And, in fact, some critics of this argument have tried to show that rule Beta is not a valid principle of inference.

In an article published shortly after Van Inwagen's *An Essay on Free Will,* David Widerker presented a first counterexample to the validity of rule Beta. The example was as follows:

> Suppose that by destroying a bit of radium r before t9, Sam prevents the emission of a subatomic particle by r at t9. Suppose further that this is the only way by which Sam can make sure that r will not emit radiation at t9. Finally suppose that Sam is the only sentient being that exists or ever existed. Let "R" and "S" stand for
>
> R: A bit of radium r emits at t9 a subatomic particle.
>
> S: Sam destroys r before t9.
>
> (Widerker 1987:38)

On the basis of this example, and applying rule Beta, we can construct the following argument:

Premise 1: N (¬R)
Premise 2: N(¬R→S)
Conclusion: N (S) rule Beta

Premise 1 is true. It is true that a certain bit of radium r does not emit a subatomic particle at t9, and it is also true that Sam does not have, or ever had, any choice about it (he could not have acted so as to ensure that r does emit the particle, so ensuring the falsity of ¬R), for, even if he does not destroy r before t9, it might still be that r does not emit the particle at t9. Premise 2 is also true. The expression "(¬R→S)" is to be taken as a material conditional (remember that, according to Van Inwagen, all the conditionals in the Consequence Argument are material conditionals); now, both ¬R and S are true, so the conditional is true; and Sam could not have acted so as to ensure its falsity; for the conditional to be false, it should be the case that the antecedent (¬R) is true and the consequent (S) is false; but, though Sam can ensure the falsity of the consequent by not destroying r at t9, he cannot ensure the truth of the antecedent; in not destroying r at t9, it might be the case that r emits the particle at t9, and then ¬R would be false as well; in this case, the conditional would be true. However, the conclusion is false, for, as we have already pointed out in commenting on the premises, Sam has a choice about the truth of S: he could have ensured the falsity of S by not destroying the bit of radium before t9.

Therefore rule Beta is invalid. It can lead from true premises to a false conclusion.

Note that, in Widerker's example, indeterminism is assumed to hold: whether the bit of radium emits a subatomic particle at t9 is not determined. We shall come back to this feature of the example. More recently, Thomas McKay and David Johnson (McKay and Johnson 1996) have produced another counterexample to rule Beta. Though similar in some respects to Widerker's, it also presents some significant differences, the main one being that it does not assume that indeterminism holds.

McKay and Johnson provide the following derivation, which uses rules Alpha and Beta:

1. N P	Premise
2. N Q	Premise
3. □(P→(Q→(P & Q)))	Necessity of logical truth
4. N(P→(Q→(P & Q)))	3, Alpha
5. N(Q→(P & Q))	1, 4, Beta
6. N(P & Q)	2, 5, Beta

So, using rules Alpha and Beta, from N P and N Q we can derive the conclusion N(P & Q). The following inference rule, which McKay and Johnson call "Agglomeration", is therefore a logical consequence of Alpha and Beta:

Agglomeration: (N P & N Q) implies N(P & Q)

Agglomeration, however, is not a valid rule. McKay and Johnson show this by means of the following example (cf. McKay and Johnson 1996:115). Suppose that I do not toss a coin, but could have done it. Let P be the proposition expressed by "the coin does not land heads", and Q the one expressed by "the coin does not land tails". In this case, both "N P" and "N Q" are true. "N P" is true, for the coin does not land heads (it does not land at all, for it is not tossed) and nobody could have ensured, by tossing it, that it lands heads. And the same holds, *mutatis mutandis,* for "N Q". Using Agglomeration, we can arrive at the conclusion that N(P & Q), that is, that the coin lands neither heads nor tails (which is true, since it is not tossed) and that nobody could ensure the falsity of this. But I could have ensured that falsity by tossing the coin. If it had landed heads, "P" would have been false and if it had landed tails, "Q" would have been false. In either case, I could have ensured the falsity of "P & Q". Therefore the conclusion of the argument, that is, "N(P & Q)", is false.

So Agglomeration is not valid. But it does follow from Alpha and Beta. Now, since Alpha is so obviously correct, McKay and Johnson conclude that Beta is not valid. Note that the example does not presuppose indeterminism. Even if, by tossing the coin, nobody can ensure in advance one particular result (heads or tails), the result that actually takes place can be perfectly determined after the coin has left the hand.

Defending incompatibilism

Though the preceding examples show conclusively that rule Beta, as it appears in Van Inwagen's Third Formal Argument, is not valid, there is still a large variety of options open to incompatibilists to defend the thesis that determinism rules out alternative possibilities, or freedom to do otherwise. One of those options is to formulate inference rules akin to, but different from, Beta, which do not succumb to counterexamples.

McKay and Johnson themselves suggest an alternative rule that can resist Widerker's example. As we said, this example assumes indeterminism. But this assumption would seem to undermine the example, for the Consequence Argument is supposed to deal with the consequences of *determinism*. As McKay and Johnson write, "assuming that the world is indeterministic is a problematic way to argue against Beta, since Beta is to be used in drawing out the consequences of determinism. If every counterexample to Beta had to be indeterministic, then a very simple revision would suffice to maintain van Inwagen's argument for incompatibilism" (McKay and Johnson 1996:118). The revision they suggest consists in replacing Beta with the following rule:

Delta: D, N P, N(P→Q) implies N Q

where "D" stands for the thesis of determinism. Widerker's example is powerless against Delta, which could then be used in an argument for incompatibilism similar to Van Inwagen's. However, McKay and Johnson hold that their own example does not assume that indeterminism is true, so that it can be used not only against Beta, but against Delta as well. We shall come back to this.

Alicia Finch and Ted Warfield proposed an additional alternative rule to Beta (Finch and Warfield 1998:521). They call it "Beta 2":

Beta 2: (N P & $\Box$(P→Q)) implies N Q

In Beta 2, the material conditional is preceded by the symbol of broad logical necessity ("$\Box$"), instead of Van Inwagen's operator N. What Beta 2 says is that "one has no choice about the logical consequences of those truths one has no choice about" (Finch and Warfield 1998:522). Now, since nobody has any choice about the past and about the natural laws, and since, given determinism, the future is a logical consequence of the past and the laws, Beta 2 allows the inference from determinism to the conclusion that nobody has freedom to do otherwise. Finch and Warfield's modified incompatibilist argument, which they call "the Improved Consequence Argument", runs as follows (where "P" stands for a proposition that expresses the complete state of the world at a time in the distant past, "L" for the conjunction of the laws of nature, and "F" for any truth):

1. $\Box$((P & L)→F)	Premise, consequence of Determinism
2. N(P & L)	Premise, fixity of the past and laws
3. N F	Conclusion, 1, 2, Beta 2

Bear in mind that, by using Beta 2, this argument circumvents McKay and Johnson's counterexample, since the invalid Agglomeration principle that they derived by means of Alpha and Beta can no longer be derived by using Alpha and Beta 2. Finch and Warfield's argument does not start from separate premises about the fixity of the past and of the laws. It rather introduces, as premise 2, the thesis that nobody has any choice about the conjunction of the past and the natural laws. But the same intuitions that back premises 4 and 6 of Van Inwagen's Third Formal Argument also support premise 2 in Finch and Warfield's incompatibilist argument. As they contend, "the conjunction (P & L) offers a description of what might be called the 'broad past'—the complete state of the world at a time in the distant past including the laws of nature. We maintain…that the broad past is fixed in just the way that Van Inwagen maintains that the past is fixed (and that the laws are fixed)" (Finch and Warfield 1998:523).

In a recent article (Huemer 2000), Michael Huemer has presented a related way of defending the incompatibility between determinism and freedom to do otherwise. Instead of replacing rule Beta as such, he proposes to change the reading of Van Inwagen's operator N. Remember that, according to Van Inwagen, "N P" is to be read as "P, and no one has, or ever had, any choice about whether P", where having a choice about whether P is to be able to act so as to *ensure* the falsity of P. This interpretation of the operator N plays an important role in the counterexamples to rule Beta that we have presented above. In McKay and Johnson's example, the agent could toss the coin, and the coin might land heads, but she cannot ensure that it does land heads. Huemer's proposal is to understand "N" in a different way, so that "NsP" should be read as follows:

NsP=No matter what S does, P

where "no matter what S does, P" is to be read as "P, and for each action, A, that S can perform, if S were to perform A, it would still be the case that P" (cf. Huemer 2000:538).

Huemer argues that, with this definition of "NsP", both rule Beta and a rule he calls Beta*, which is in fact McKay and Johnson's "Agglomeration", are valid. If "no matter what S does, P" and "no matter what S does, Q" are both true, then it seems that "no matter what S does, P and Q" will also be true. Huemer's proposal amounts, in fact, to replacing rule Beta by an alternative rule, in that Van Inwagen's operator N, which appears in his formulation of Beta, is actually replaced by a different operator.

It is clear that, if Huemer's proposal is correct, Widerker's is not a counterexample to a version of the Consequence Argument in which the operator N receives the new interpretation. In Van Inwagen's reading of the operator, "N P" is true in Widerker's example: the bit of radium r does not emit a subatomic particle at t9 and Sam has no choice about this, for, even if he does not destroy r before t9, it still might be that r does not emit the particle at t9. However, in Huemer's reading, "N P" is not true. It is not the case that, no matter what Sam does, r does not emit the particle at t9, for it is open to him not to destroy r before t9 and, if he does not destroy r before t9, it might be the case that r emits the particle at t9. In fact, in his paper Huemer himself presents counterexamples to Van Inwagen's rule Beta that have a similar structure to Widerker's and that also presuppose indeterminism.

So even if counterexamples to Van Inwagen's rule Beta in which indeterminism is assumed show that this rule is not valid, they are not effective against the incompatibility between determinism and alternative possibilities, for rule Beta can be replaced by, e.g., Finch and Warfield's Beta 2, or McKay and Johnson's Delta, or Huemer's new rule Beta, and new incompatibilist arguments can be devised that are not threatened by examples of that sort. A general lesson to be drawn from this is, as Thomas Crisp and Ted Warfield put it in a recent article, that "proposed counterexamples to Principle Beta must not presuppose the truth of indeterminism" (Crisp and Warfield 2000:180).

However, McKay and Johnson claim, as we saw, that their example does not presuppose the truth of indeterminism. Coin tossing might be a deterministic process, even if emission of subatomic particles were not. Unlike Widerker, McKay and Johnson do not assume that the world in which their example of the coin toss takes place is indeterministic. In fact, they would do better to assume that it is not, for otherwise their example would be powerless against their own rule Delta (and against such proposed inference rules as Beta 2 or Huemer's new rule Beta). Suppose, then, that their example takes place in a deterministic world. According to this supposition, McKay and Johnson's example fails as a counterexample to Van Inwagen's rule Beta, as Crisp and Warfield (2000) have convincingly shown. Their point could be put as follows. In this (deterministic) world, if I tossed the coin, the past and the natural laws, together with my tossing, would imply a determinate result (either heads or tails). Remember that I do not in fact toss the coin. So it is true that P (the coin does not land heads) and it is true that Q (the coin does not land tails), because the coin does not land at all. Now, if I were to toss the coin, and if determinism holds, either P or Q would be a logical consequence of my action, together with the past and the natural laws. But then, as Crisp and Warfield write: "Either it's in my power to take an action such that the action's occurrence together with the past and laws of nature ensures that ¬p, or it's in my power to take an action such that the action's occurrence together with the past and laws of nature ensures that ¬q. If the former, then Np is false; if the latter, then Nq is false" (Crisp and Warfield 2000:182). But McKay and Johnson need the truth of both "Np" and "Nq", for they argue

that, in their example, they are both true while "N(p & q)" is false. And this is what they cannot have if determinism holds.

Let us add that this problem, which arises from the assumption of determinism in McKay and Johnson's example, can be seen even more clearly if we read "N" as in Huemer's proposal. Suppose that, if I toss the coin, P is what follows from this action, together with the past and the natural laws. Then "N P" would be true, for, whether or not I toss the coin, it does not land heads. But "N Q" would be false for, though the coin actually does not land tails, if I tossed it, it would land tails. And the opposite would hold if Q, instead of P, were the logical consequence of the past and the natural laws.

Though this would seem to be a decisive response to McKay and Johnson, Crisp and Warfield have also argued against them on another basis. Since the Consequence Argument is an argument against the compatibility of determinism and freedom to do otherwise, examples that presuppose that this compatibility holds are clearly flawed. They put this in the form of a desideratum: "Proposed counterexamples to Principle Beta must not presuppose the compatibility of freedom and determinism" (Crisp and Warfield 2000:175). Now, this is actually the case in McKay and Johnson's example given the assumption that determinism holds. For they introduce this example as follows: "Suppose that I do not toss a coin but could have" (McKay and Johnson 1996:118). And this, if the world is deterministic, is to assume that determinism is compatible with alternative possibilities of action.

The upshot of all this is that, even if the main counterexamples to Van Inwagen's rule Beta that have been proposed up to now show that the rule is not formally valid, they do not seriously threaten the thesis of the incompatibility between determinism and freedom to do otherwise. There are too many ways in which rule Beta can be replaced and too many ways in which the Consequence Argument can be modified for incompatibilists to feel that the incompatibility thesis is really in trouble. In the next section, we shall see still other ways in which this thesis can be argued for.

Incompatibilism without transfer

One additional option open to defenders of the incompatibility of determinism and freedom to do otherwise is to argue for this thesis without using a rule or Principle of Transfer akin to Van Inwagen's Beta. Van Inwagen himself does not think this is a promising way. He writes: "I do not know how to prove this, but I would suppose that what is *in effect* an allegiance to rule Beta must lurk somewhere, in however inarticulate a form, in the background of any technically satisfactory argument for incompatibilism" (Van Inwagen 1994:98). But some thinkers do not agree with this. John Martin Fischer and Mark Ravizza, for example, have proposed an argument that, they hold, does not rely on rule Beta or any other Transfer Principle. It starts from the Principle of the Fixity of the Past and Laws, according to which, in Fischer and Ravizza's own words, "…an agent has it within his power to do A only if his doing A can be an extension of the actual past, holding the natural laws fixed" (Fischer and Ravizza 1998:22). On the basis of this principle, and assuming the truth of causal determinism, their argument runs as follows:

Suppose…that someone S does A at time T3. It follows from the truth of causal determinism that the state of the world at T1 together with the natural laws *entails* that S does A at T3… Given the entailment just described, S's refraining from doing A at T3 *cannot* be an extension of the actual past, holding the laws of nature fixed…S cannot at T2 refrain from doing A at T3 (…T2 is prior or contemporaneous with T3). That is to say, given the truth of causal determinism, it follows that S cannot do other than he actually does…

(Fischer and Ravizza 1998:22)

It is doubtful whether this argument does not actually rely implicitly on a rule akin to Beta. In fact, that someone "cannot do other than he actually does" does not follow from "the truth of causal determinism" alone, but from this together with the Principle of the Fixity of the Past and Laws. And the question is what licenses the incompatibilist conclusion from these two premises. Remember that this principle and the thesis of determinism were also the two premises in Finch and Warfield's argument for incompatibilism which was presented in the preceding section, but the incompatibilist conclusion from these two premises was reached there by means of a Transfer rule, namely Beta 2. And it is an open question whether the conclusion of Fischer and Ravizza's argument is not also reached by an implicit application of a Transfer rule.

Fischer and Ravizza's, however, is only one among several other attempts to argue for incompatibilism without the aid of Transfer rules. Another interesting proposal on these lines can be found in a recent article by Ted Warfield (cf. Warfield 2000). And, even more recently, Peter Unger has argued for the incompatibility of determinism and freedom to do otherwise on the basis of, as he puts it, "a line of thinking so perfectly simple and, I think, so obviously correct [that] it should hardly be called a 'philosophical argument'" (Unger 2002:5). According to Unger, the core of determinism is Inevitabilism, according to which "just as it is with the past, so the *future* is absolutely settled and closed, in every real respect and regard" (Unger 2002:3). Now, this is Unger's proposal:

Let's suppose that, as regards anything that happens after a certain time before I ever existed, at least from that time onward it is absolutely inevitable that the thing happen. Then, for each time throughout my existence—and forever after, there's really *just one* (perfectly specific) way for the world then to be. But, for any such time, I will have available alternatives, as regards what to do, only if there are *at least two different* ways for the world to be at that very moment or, perhaps, at the very next moment… So, throughout my existence whatever happens is so inevitable that I never have any actually available alternatives as regards what I do.

(Unger 2002:5)

Therefore, if determinism holds, nobody has, in Unger's words, "any full choice, or free will". What Unger seems to be doing in this defence of the incompatibility thesis is just to derive this thesis out of the very concept of determinism itself, without the aid of a Transfer rule.

Whatever the merits of Unger's argument, there would seem to be an even simpler way of arguing for the incompatibility thesis on the basis of the concept of determinism. Let us try this path. Remember that one way in which Van Inwagen defined determinism was the following: "*Determinism*…is the thesis that there is at any instant exactly one physically possible future" (Van Inwagen 1983:3). Now, think of an action I actually perform at a particular time t_2, such as reading Chapter Eight of Iris Murdoch's *The Nice and the Good*. That I am doing this at t_2 is part of the *future* with respect to an arbitrarily chosen instant in the past, t_1. Suppose that, when I perform that action, I could be performing an alternative action instead. My performing this alternative action would be part of a different future with respect to t_1. So, if I could perform this alternative action at t_2, then there would be more than one physically possible future at instant t_1. But this contradicts the definition of determinism with which we started. Therefore, if determinism holds, I could not do otherwise than read Chapter Eight of Murdoch's novel at t_2. And, since there is nothing special about me, about my reading that chapter of Murdoch's novel, or about the time (t_2) at which I do that, the incompatibility thesis follows: if determinism is true, nobody could ever have done otherwise than she in fact does.

Though Fischer and Ravizza's, Unger's or my own argument do not explicitly rely on some Transfer rule, it is difficult to ascertain whether they *implicitly* make use of one. But even if they did, incompatibilists have plenty of options, as we have seen, to maintain their position against criticisms of Van Inwagen's rule Beta and to construct versions of the Consequence argument that are not affected by those criticisms.

Conclusion

In this chapter, we have argued for premise A of the sceptical argument about moral responsibility (SMR) and we have also started to argue for premise B of that argument, namely that, if determinism is true, moral responsibility is not possible. We have argued for this by arguing in favour of premise 2 of the Incompatibilist Argument, according to which, if determinism is true, nobody could have done otherwise than she in fact did. Though a conclusive proof is hardly to be expected with respect to any important philosophical thesis,[2] we certainly have, on the basis of the considerations in this chapter, a very strong case in favour of this premise 2. However, the incompatibility between determinism and moral responsibility (SMR's premise B) does not follow from this premise 2 alone, but from this and the thesis that alternative possibilities are necessary for moral responsibility. This thesis is premise 1 of the Incompatibilist Argument. Some thinkers who accept that determinism is incompatible with alternative possibilities are not thereby convinced that determinism is incompatible with moral responsibility, for they think that alternative possibilities are not actually required for moral responsibility. They deny thus premise 1 of the Incompatibilist Argument. If their main concern is about moral responsibility, they will not think that discussions about the Consequence

Argument, or about other arguments to the effect that determinism is incompatible with alternative possibilities, are actually very pressing for them. So our next task will be to discuss premise 1 of the Incompatibilist Argument, according to which moral responsibility requires alternative possibilities, so that an agent is morally responsible for an action of hers only if she could have done otherwise. This contention will be the subject of the next chapter.

2
Alternative possibilities and moral responsibility
(SMR's premise B)

Remember the Incompatibilist Argument in favour of SMR's premise B which was presented in the previous chapter? It goes: 1) Moral responsibility requires alternative possibilities: an agent is morally responsible for an action of hers only if she could have done otherwise. 2) Determinism rules out alternative possibilities: if determinism is true, nobody could have done otherwise than she in fact did. The conclusion of these two premises is SMR's premise B, namely that, if determinism is true, moral responsibility is not possible.

SMR's premise B asserts the incompatibility between determinism and moral responsibility. Compatibilists deny this thesis. They can resist it by rejecting either of the premises of the Incompatibilist Argument (or both). In the preceding chapter we discussed objections to premise 2, and argued that they are not successful. In this chapter, we shall present and discuss some ways in which premise 1 could be questioned.

Frankfurt cases

More than thirty years ago, in a celebrated and much discussed article (Frankfurt 1969), Harry Frankfurt made a strong case against the assumption that moral responsibility requires alternative possibilities (premise 1 above), an assumption which he calls "the principle of alternate possibilities". According to this principle, in Frankfurt's own words, "a person is morally responsible for what he has done only if he could have done otherwise" (Frankfurt 1969:1). In spite of its general and virtually unquestioned acceptance, as well as its initial plausibility, which has led some philosophers to think of it as an a priori truth, Frankfurt contends that the principle is false and that its plausibility is only an illusory appearance. Before going into Frankfurt's criticism, let us stress that the principle of alternate possibilities (PAP) plays a central role in our intuitions about moral responsibility, which may partly explain why, for a long time, nobody has been inclined to deny it. In addition to its role in the dispute about the compatibility of determinism and moral responsibility, PAP is also important because it affects a reasonable desire about the moral responsibility we are prepared to bear for what we do. Understandably enough, we want to have *control* over the (degree of) moral responsibility we bear for our actions. And that is partly why we want PAP to be true: we want it to be true that, if we are to be morally responsible for a certain action, we have freedom to do that action or to do something else instead, including simply not doing that action.[1] If this is false, it seems that we lose that control, for then we can be morally responsible for actions that we could not avoid performing. In relation to this, there also

seems to be a logical connection between PAP and another venerable moral principle, namely that "ought" implies "can" (OIC). It has recently been argued that OIC, together with some plausible assumptions, implies PAP, so that rejecting PAP may imply rejecting OIC as well (cf. Widerker 1991 and Schnall 2001).[2] One way of seeing these connections, in the case of moral blameworthiness, is as follows. When we blame someone for something she did, A, we do it under the assumption that she ought (she was morally obliged) not to have A-ed. Now, following OIC, this assumption implies that she was able not to A. Suppose, however, that she was not able not to A, so that her A-ing was inevitable. Then, by OIC, she was not morally obliged not to A. But this means that our assumption fails and that our judgement that she was blameworthy for A-ing is unjustified, or simply false. So, on the basis of OIC and of the assumption that moral blameworthiness for A-ing requires moral obligation not to A, we can go from the lack of alternative possibilities with respect to A to the lack of blameworthiness for A-ing, which is logically equivalent to PAP as applied to moral blameworthiness. All this looks extremely plausible.[3] Denying it means to find it acceptable to burden people with moral duties that they cannot discharge, and to hold them morally responsible for actions that it is not in their power not to perform. Too much moral weight to carry, it seems. So, if rejecting PAP actually implies rejecting OIC and accepting all this moral weight on one's shoulders, one should think twice before dropping that principle.

According to Frankfurt, however, PAP is actually false. Its apparent plausibility derives from a confusion of this principle with the true assumption according to which coercion, beyond a certain degree, diminishes or even rules out an agent's moral responsibility. It might be thought that this contention is a particular case of PAP, so that the truth of this contention derives from the fact that the coerced person cannot do otherwise (cf. Frankfurt 1969:2). But this is not right, Frankfurt contends. What rules out the agent's moral responsibility in cases of coercion is not her lack of alternatives. This can be seen by considering cases in which a person does something within circumstances that leave no alternative to doing it, but that do not cause nor causally explain her actually doing this. In cases like this, our judgement is that the agent may bear full responsibility for her action, despite her lack of alternative possibilities. Frankfurt starts by considering cases where both the agent's decision to do A and a coercion or threat to do A are present. In cases like this, if the agent acts because of her own decision, and not because of the coercion, she can be morally responsible for her action, though, in the circumstances, she could not have avoided performing it. In fact, coercion excludes moral responsibility only when it accounts for the agent's action, and not by virtue of the mere fact that it excludes alternative possibilities.

Frankfurt considers a likely objection a defender of PAP will raise. According to this objection, cases of this sort do not show PAP to be false, for it is still open to the agent to defy the threat she knows to be present and accept the consequences, so that she has, after all, alternative possibilities and PAP has not been refuted.

Another objection, which Frankfurt does not actually consider, is that, in the case at hand, our intuitions about what does actually cause the agent to act as she does may be unsteady. Given that the agent is, after all, aware of the threat, her decision might not be the sole cause of her action. Maybe the threat, perhaps half-consciously, was also playing a causal role in her taking that decision. If so, this would be a case of coercion, and we

might withdraw our judgement that the agent is fully responsible for what she did. Without this judgement, however, PAP is not undermined by the example.

However, though Frankfurt does not consider this latter objection, he does in fact meet it, as well as the one he actually considers. He does so by constructing an example in which the factor that ensures the agent's lack of alternatives is (unlike the threat in the previous examples) merely counterfactual and completely unknown to the agent himself. He writes:

> Suppose someone—Black, let us say—wants Jones to perform a certain action. Black is prepared to go to considerable lengths to get his way, but he prefers to avoid showing his hand unnecessarily. So he waits until Jones is about to make up his mind about what to do, and he does nothing unless it is clear to him (Black is an excellent judge of such things) that Jones is going to decide to do something *other* than what he wants him to do. If it does become clear that Jones is going to decide to do something else, Black takes effective steps to ensure that Jones decides to do, and that he does do, what he wants him to do. Whatever Jones's initial preferences and inclinations, then, Black will have his way.
>
> (Frankfurt 1969:6. Index to "Jones" deleted)

Suppose, however, that Black has no need to intervene, because Jones, on the basis of his own motives and preferences, and reasoning fully on his own, decides to do, and does, precisely what Black wants him to do. Given Black's lurking presence, Jones could not have done otherwise and, none the less, he seems to bear exactly the same moral responsibility and deserve exactly the same praise or blame as he would if Black had been fully absent. Jones's lack of alternatives "played no role at all in leading him to act as he did" (Frankfurt 1969:7).

Let us assume, as is commonly done (though this addition is not Frankfurt's, but, as far as I know, Dennett's), that Jones's action is killing Smith and that Black, without Jones's awareness, has implanted in Jones's brain a device that monitors Jones's mental processes and deliberations without interfering with them unless Black presses a special button. Of course, Black does not need to press the button, since Jones deliberates and decides to kill Smith of his own accord. What Frankfurt intends, as I understand him, is to test our intuitions about the necessity of alternative possibilities for moral responsibility, by casting light on some assumptions we actually, though perhaps unknowingly, bring into play in our moral responsibility ascriptions. If, in a case in which alternative possibilities have been artificially removed, with this fact playing no causal or explanatory role in the agent's decision and action, we definitely judge that she is morally responsible for what she does, it seems clear, surprising as this result may appear to us, that an agent can be morally responsible for what she did even if she could not have done otherwise, and PAP is false. Moreover, moral responsibility would be compatible with the lack of alternative possibilities not only of action but even of decision, for note that Black's intervention, which never takes place, might occur when Jones "was about to make up his mind"; that is, before a full decision was made by him.

Is PAP actually refuted by Frankfurt's example? To anticipate, our considered answer will be that it is not. But justifying this answer is a complex task. Frankfurt's argument has given rise to a huge literature and to a large and subtle discussion, which is still going

on. Reviewing the entire discussion would be an almost endless task, but we shall try to draw its main lines.

Frankfurt's contention is strongly dependent on our having pretty clear intuitions to the effect that Jones is morally responsible for what he did. If these intuitions are dubious or simply absent, the argument fails. As a matter of fact, most participants in the discussion, with either compatibilist or incompatibilist inclinations, accept Frankfurt's judgement that Jones is morally responsible for his action. I side with this majority party. These intuitions are plausibly explained by our feeling that Jones would have deliberated, decided and acted in the same way even if Black had not been there. Black is a purely counterfactual intervenes He did not have any causal influence on Jones's deliberation, decision and action. If one tries to face this example leaving aside any prior commitment to incompatibilism, then, on the natural and charitable assumption that no common ground responsibility-undermining circumstances, such as coercion, pathological compulsion or a seriously impaired capacity for decision making, obtain, the judgement that Jones is responsible is quite natural and even compelling. This can be seen in the following nice text of Van Inwagen's (with Gunnar, Cosser and Ridley substituted for Jones, Black and Smith respectively):

Let us suppose, for the moment, that Cosser had not devised his plan for ensuring that Gunnar shoot Ridley. Let us consider Gunnar's wicked act and let us "build into" our description of the circumstances under which it was performed *whatever* may be necessary for Gunnar's being responsible for it… Now let us consider adding to this description of the circumstances of Gunnar's act the statement that Cosser *would have* caused him to perform it if he had changed his mind. Does this statement alter the fact of Gunnar's responsibility? How could it?… The causal history of his act is just what it would have been if Cosser had never existed.

(Van Inwagen 1983:163–4)

However, accepting that Jones is responsible in Frankfurt's case does not by itself entail accepting the falsity of PAP (or at least of some closely related principle). It is open to defenders of PAP (or a closely related principle) to insist that, though Jones was indeed responsible for what he did in the actual sequence, he none the less had alternatives of some kind that can ground his moral responsibility. This resistance move against Frankfurt's attack on PAP has a rather long history. Let us go into it.

Searching for alternatives

The possibility of defending PAP against Frankfurt's attack by finding alternatives open to Jones was in fact suggested by Frankfurt himself in his original paper, for he acknowledges that "it is in a way up to him whether he [Jones] acts on his own or as a result of Black's intervention" (Frankfurt 1969:8). Moreover, the very structure of Frankfurt's example (and others in the same style) requires that the agent have alternatives of some sort. The reason is that Black's intervention is counterfactual and contingent upon Jones's showing some sign that he is going to make the "wrong" (in Black's eyes) decision. But then it must be true of Jones that, even if he shows no

such sign (as he in fact does not), he *could* have shown it. So Frankfurt's example paves the way for the strategy we are considering.

However, discovering alternatives is clearly not enough to resist Frankfurt-style arguments. It also has to be shown that these alternative ways, which the agent did not take, were relevant to his moral responsibility in the actual sequence. This remark is important, for not all alternatives that are open to an agent are relevant to her moral responsibility for a certain action. To take an extreme example, which I borrow from Pereboom, imagine an entirely deterministic world, with no alternative possibilities whatsoever, where a person decides on her own to evade taxes and does so; suppose, however, that ten years ago God made a miracle, giving rise to a short indeterministic period in the agent's life and allowing her to effectively choose between, say, having potatoes or beans for dinner; she chose potatoes, but choosing beans was also available to her; shortly after that, the world returned to its deterministic state; now, how could the agent's having these alternatives possibly have any bearing on her moral responsibility, if she has it, for evading taxes?

In fact, a touchstone for attempts to resist Frankfurt's attack on PAP on the basis of finding alternatives open to the agent will be the relevance of those alternatives to the responsibility attribution. The agent's having these alternative possibilities should at least partly explain why she bears responsibility for what she in fact decided and did, or how the fact that she has these alternatives justifies (or improves the justification of) our judgement that she is responsible. Actually, the main objection to this line of resistance, which Fischer has called, somewhat disparagingly, the "flicker of freedom" strategy, will be the "robustness" objection (the term is also Fischer's), which points to the lack of relevance of the proposed alternative possibilities for the agent's moral responsibility.

Let us present some versions of this sort of attempt to defend the necessity of alternative possibilities for moral responsibility against Frankfurt-style arguments.

A rather sophisticated version of this strategy has been developed by Van Inwagen (Van Inwagen 1983:161–80). As a preliminary point in his defence of freedom to do otherwise as a requirement for moral responsibility, Van Inwagen stresses the connection, which we indicated earlier, between that freedom (which PAP gives an expression to) and the principle that "ought" implies "can". As we saw in the preceding section, Van Inwagen readily accepts Gunnar's (Jones's in Frankfurt's example) moral responsibility for killing Ridley (Smith) and he is even prepared to allow that, as a consequence of this, PAP may be false. He thinks, however, that some principles, akin to PAP, are true and not affected by Frankfurt-style counterexamples, and that they still show that alternative possibilities are required for moral responsibility. The first of these principles is the Principle of Possible Action (PPA), according to which "a person is morally responsible for failing to perform a given act only if he could have performed that act" (Van Inwagen 1983:165). So, even if I decide on my own not to do something, I am not responsible for not doing it if, in fact, I could not have done it. If I do not telephone the police to prevent a robbery I am witnessing but, unknown to me, there is then a breakdown in the telephone network, I am not responsible for not calling the police, for I could not have called them. What I am responsible for is perhaps not *trying* to call the police, which I could have done.

Both PAP and PPA are principles about acts, performed or unperformed, but, as Van Inwagen says, ascriptions of moral responsibility relate more commonly to results or consequences of these acts or failures to act, to events or states of affairs. In Frankfurt's example, for instance, it would be typical to say of Jones that he is responsible for Smith's death rather than for killing him. Van Inwagen goes on to state principles about moral responsibility for events and states of affairs. Since it is not clear whether these are particulars or universals, he states two separate principles, one for particular events or states of affairs and the other for events or states of affairs as belonging to general kinds. He calls these principles "principles of possible prevention". The first (PPP1), which applies to particular events, is as follows: "A person is morally responsible for a certain event-particular only if he could have prevented it" (Van Inwagen 1983:167). The second (PPP2) is about universals: "A person is morally responsible for a certain state of affairs only if (that state of affairs obtains and) he could have prevented it from obtaining" (Van Inwagen 1983:171).

In order to apply PPP1 to concrete cases of moral responsibility for a certain particular event, we need individuation criteria for events: "We want to know how to tell of some given event whether *it,* that very same event, would nevertheless have happened if things had been different in certain specified ways" (Van Inwagen 1983:168). In Frankfurt's example, we want to know, of the particular event that Jones brought about in the actual sequence, whether *it* would also have occurred if Black had intervened. Van Inwagen's proposal is to use a truncated version of Davidson's criterion of event individuation. According to this version, A and B, where "A" and "B" denote particular events, are one and the same event if, and only if, they have the same causes. Suppose, for example, that Cleopatra poisoned Caesar, so causing his death. This event, which it would have been right to call "Caesar's death", would not have been the same particular event as Caesar's death as it actually happened, since the latter was brought about by different causes.

We can see now that an expression such as "Caesar's death" is ambiguous. It can refer to a particular event, e.g. to Caesar's death as it was brought about as a matter of actual fact, or to a kind of event or event-universal, say the fact that Caesar dies, which can be exemplified in different ways and have different causes. So both Caesar's actual death and his death had Cleopatra poisoned him would be exemplifications of the same kind of event or event-universal, but they are different particular events.

Armed with these distinctions, we can now see what follows for Frankfurt's example. It invites us to draw the conclusion that Jones is responsible for Smith's death, even if he could not have prevented it from occurring, so apparently contradicting PPP1. But we can now see that this conclusion is too rash. If "Smith's death" consistently names a particular event all along, Jones is morally responsible for it, but he *could* have prevented it from occurring. If he had refrained or showed clear signs that he was not going to kill Smith, then Black would have intervened and Jones would have brought about Smith's death. But this would have been a different particular event, since it would have had different causes. So Jones was responsible for Smith's death but had alternatives. PPP2 is not violated either. If "Smith's death" refers to an event-universal, then Jones could not have prevented it from occurring, for Smith would have died one way or another, but Jones is not responsible for *that* either, which is what follows from PPP2. So there is no consistent interpretation of "Smith's death" in which it is true both that Jones is

responsible for it and that he could not have prevented it. The necessity of alternatives for moral responsibility is thus vindicated against Frankfurt's challenge.

With some hesitations, Van Inwagen thinks that PAP cannot be vindicated by having recourse to a corresponding distinction between "act-particulars" and "act-universals". He closes this path for himself by interpreting act-particulars as "event-particulars that are voluntary movements of human bodies" (Van Inwagen 1983:180). But I see no reason why one should think of particular acts in such a restricted way. We do have the intuitive distinction between particular acts and kinds of acts. "Evading taxes" is a kind of act, which can be (and actually is) exemplified by different particular persons at different places and times. Van Inwagen is right in saying that we do not usually ascribe responsibility for voluntary movements of human bodies. But a person can be morally responsible for a particular act of evading taxes performed by her, and this does not reduce to a voluntary movement of her body. Moreover, we have theoretical elaborations of this intuitive distinction. Davidson's theory of action (Davidson 1982), for instance, relies heavily on a closely related distinction between action-tokens and action-types. The notion of a particular action being described in different ways and so exemplifying several action-types or act-universals is central to a Davidsonian approach to the philosophy of action. We certainly need criteria of individuation for particular acts, but these do not seem to face special difficulties as compared with individuation criteria for particular non-actional events. In Frankfurt's example, we see no reason why we could not distinguish between the particular act performed by Jones and several act-universals that this particular act exemplifies. If so, then PAP itself, and not only other related principles, could be defended against Frankfurt's challenge by showing that his is not really a case in which a person is responsible for something (act-particular or act-universal) while having no alternative to *that*. We shall develop this line of argument below, in trying to respond to Fischer's, Pereboom's and Zagzebski's objections to the line of resistance against Frankfurt's attack on PAP which we are now considering.

A second version of this line can be found in Margery Bedford Naylor (Naylor 1984) and Donald Davidson (Davidson 1973), among others. This version relies on a careful specification of what an agent is responsible for, in order to show that she has alternative possibilities, even in Frankfurt's or Frankfurt-type examples. If we look carefully at these examples, we can see that there are differences between what the agent does in the actual sequence and what she would do in the alternative sequence, if the counterfactual intervener were to interfere in the process. In Frankfurt's example, though Jones kills Smith both in the actual and in the counterfactual sequence, he only does it *on his own* in the former sequence, for, when Black intervenes, Jones does not kill Smith on his own, but, say, as an instrument in Black's hands. So, Naylor writes: "Since it was entirely up to Jones whether or not to do A on his own, he is clearly responsible for doing A on his own. But it is not obvious that he is morally responsible for doing A" (Naylor 1984:251). In fact, her final conclusion supports the connection between moral responsibility and alternative possibilities: "Jones is morally responsible for doing A on his own because it was within his power not to do A on his own; but he is *not* morally responsible for doing A because it was *not* within his power not to do A" (Naylor 1984:257). If what we hold Jones to be responsible for is doing A on his own, then Jones has alternatives, and PAP is vindicated.

Though Davidson deals with this issue more briefly, I think his position can be rightly classified within this second version of the general strategy we are considering. According to Davidson, what depends on the agent, in Frankfurt cases, is not properly whether he does A, but whether he does A *intentionally*. He writes:

> It is true that it may sometimes be the case that what a man does intentionally he might have been caused to do anyway, by alien forces. But in that case what he would have done would not have been intentional. So even in the overdetermined cases, something rests with the agent. Not, as it happens, *what* he does (when described in a way that leaves it open whether it was intentional), but whether he does it intentionally. His action, in the sense in which action depends on intentionality, occurs or not as he wills; what he does, in the broader sense, may occur whether or not he wills it.
>
> (Davidson 1973:150)

So, both for Naylor and for Davidson, the agent, even in Frankfurt cases, where he is overdetermined to do A (by his own decision and, if this fails, by an alien intervention), has alternatives: it is within his power to do A on his own or not, and to do A intentionally or not. Unlike Naylor, Davidson does not explicitly address the issue of the agent's responsibility, but it is plausible to think he would concur with Naylor on this account.

A third version of the strategy of looking for alternatives in Frankfurt cases, in order to save the connection between alternatives and moral responsibility, focuses on the event which triggers the intervention of the counterfactual factor. In Fischer's terms, while the two versions we have considered are forward-looking, this third version looks backwards, searching in the actual sequence for a sign that Jones *could* have shown (but did not) and that would have triggered Black's intervention. Unless determinism is assumed, it seems clear that it must rest with the agent *to show that sign or not,* so that he has at least these alternative possibilities. Even if deciding to do otherwise is not open to the agent, given that, in Frankfurt's original example, Black's counterfactual intervention takes place before Jones's decision, Jones can have open to him such pathways as trying to decide not to kill Smith, or forming a preference for, or showing an inclination towards, that alternative way of acting. In fact, commenting on this version of what he calls the "flicker of freedom strategy", Fischer writes that "it is hard to see how a Frankfurt-type example could be constructed which would have absolutely *no* such flicker. For a Frankfurt-type case must have an alternative sequence in which intervention is triggered in some fashion or other… Thus, it appears that, no matter how sophisticated the Frankfurt-type example, if one traces 'backward'…far enough, one will find a flicker of freedom" (Fischer 1994:136).

Despite the ingenuity displayed in finding alternatives even in Frankfurt cases, this strategy faces some important difficulties. Let us now look at some of them.

Criticisms of the preceding strategy: the "robustness" objection

Consider first Van Inwagen's proposal. One worry that one can have concerns the individuation criterion for particular events. It does not seem clear that, just by virtue of having different causes, a particular event is *ipso facto* different from another. Consider

the event that is my actually purchasing a newspaper at a particular time and at a particular newsagent's. Suppose I buy the newspaper because I want to know the latest sports news. Would this particular event of my buying the newspaper not have occurred if I had bought the same newspaper at the same time and place just because I wanted to know, say, the latest political news? This looks implausible, at least. If it is, then Van Inwagen's strategy against Frankfurt may be undermined, for now the door is open for the possibility that, in Frankfurt's example, Jones may bring about the same particular event in the alternative sequence as in the actual sequence, in spite of their different causes. Suppose that, in the alternative sequence, where Black intervenes, Jones causes Smith's death at the same moment and place, and with the same bodily movements as in the actual sequence. It is not obvious that, in this case, the particular event of Smith's death does not take place both in the actual and in the counterfactual sequence. But if it does, then PPP1 would be shown to be false, for Jones would be responsible for the particular event which was Smith's death even if he could not have prevented that particular event from occurring.

In my view, however, Van Inwagen's proposal can be elaborated and made stronger provided that the difference between what happens in the actual sequence and in the alternative sequence is made intuitively obvious, instead of relying on individuation criteria that may be contentious. And a way of achieving this is to extend his proposal to actions, as we have suggested it could be. We shall put forward this line of argument below, in the context of a general response to Frankfurt-type objections to the necessity of alternatives for moral responsibility.

Naylor's and Davidson's versions of the strategy also face some problems. According to Naylor's version, Jones is morally responsible for doing A *on his own,* but to this he had alternatives; however, he is not morally responsible for doing A, for he did not have alternatives to this. But, as Kane points out, Naylor's position "too artificially separates responsibility for doing-A-on-one's-own from responsibility for doing A. In general, if we are responsible for doing something on our own, we are responsible for doing it" (Kane 1996:41).

This criticism seems fair enough. If Davidson's position is developed in the same vein as Naylor's, it will also face this problem, substituting "intentionally" for "on one's own" in the Kane quotation.

But the most important criticism to this strategy, in all the versions we have considered, has been raised by Fischer (1994), and endorsed and elaborated further by other thinkers, especially Derk Pereboom (2000, 2001, 2003) and Linda Zagzebski (2000a). We have already anticipated it in insisting that not all alternatives a subject has are relevant to her moral responsibility for a given action or a consequence thereof. Following Fischer, we may dub this criticism "the robustness objection". Let us see how it goes.

We may start by pointing out that, when we think about alternative possibilities in relation to moral responsibility attributions, the most natural way of seeing the matter is as follows. We imagine the agent confronting two (or more) alternative actions, deliberating about them and deciding which one she will perform, so that, though she chooses and does, say, A, she could also have freely chosen and done B instead. That is, we ordinarily suppose that the alternative that the agent does not choose is one that she was free to have chosen and acted upon. In this context, if we judge that the agent chose

and did the wrong (right) thing, we find her blameworthy (praiseworthy), at least in part, because we assume that it was up to her (she was free) to choose and do the other. If we find her blameworthy, it is partly because we assume that she ought, and it was open to her, not to have acted that way. If we find her praiseworthy, it is partly because we assume that she did what she ought to do and that it was up to her to have done otherwise. It is in this sense that having alternatives is (explanatorily) relevant for moral responsibility and for attributions thereof. Let us call alternatives that meet this criterion of explanatory relevance "robust alternatives". Characteristically, decisions that the agent was free to make and actions that she was free to perform are robust in this sense.

Now, the robustness objection, which we owe to Fischer, claims that even if, in Frankfurt cases, the agent has alternatives of some sort, these are not robust enough; that is, they are not the sort of alternatives that can ground moral responsibility ascriptions. So, even if the agent can give rise to a different particular event, or act *not* on her own, as the first and second versions of the criticised strategy respectively have it, these alternatives are not of the right kind. As Fischer points out, in relation to the first, "it is highly implausible to suppose that it is *in virtue* of the existence of such an alternative that Jones is morally responsible for what he does" (Fischer 1994:140).

A crucial reason for this contention is that, in Frankfurt cases, the alternative pathways, where Black takes over, are such that the agent *does not act freely* along them. Alternative possibility views conceive of the agent as having *control* over which alternative possibility becomes actual, and it is precisely by virtue of having this sort of control, which Fischer calls "regulative control", that the agent can be legitimately held responsible for what she actually decides and does. That the agent has this sort of control is what we imply when we say that she was *free* not to have done what she actually did, and this is (at least partly) why we judge that she is morally responsible for that. But, as Fischer writes:

> How exactly *could* the existence of various alternative pathways along which the agent does *not* act freely render it true that the agent has the relevant kind of control (regulative control)? And notice that this is precisely the situation in the Frankfurt-type cases (…) It is not… plausible to suggest that it is in virtue of a set of alternative possibilities in which Jones does *not* act freely that he actually can be held morally responsible for his behavior. How could adding a set of alternatives in which Jones does *not* act freely make it the case that he *actually* acts freely?
>
> (Fischer 1994:141, 142)

The "flicker of freedom" theorist wants to obtain freedom and responsibility in the actual sequence from an alternative sequence where freedom and responsibility are absent. And this would seem to involve a sort of mysterious *alchemy,* in Fischer's own terms.

Even if we agree with Van Inwagen that, in Frankfurt's case, Jones brings about a different particular event in the alternative sequence than in the actual sequence, he does not choose or do it freely, owing to Black's intervention. Equally, if we agree with Naylor that, in the alternative sequence, unlike the actual one, Jones does not kill Smith on his own, this is not something that he freely chooses and does. So Fischer's objection seems clearly to apply to these two, forward-looking versions of the criticised strategy.

However, the third, backward-looking version of the strategy, according to which the agent, even if she cannot strictly decide or do otherwise, can none the less try to decide otherwise, or begin to do so, or form a different preference, would seem to have a resistance movement at its disposal. Granting that the alternatives identified look rather slender, it is open to defenders of this version to hold that those alternative things are none the less *freely done* by the agent. Consider that the alternatives mentioned can be seen as action-like and so, in some sense, as subject to the agent's will. Of course, this spark of freedom quickly goes out owing to Black's intervention and, thereafter, the subject decides and acts unfreely. Since these typically robust alternatives, namely decisions and actions, are not freely performed, it is dubious whether such slender, subtle alternatives as remain are robust enough to ground moral responsibility ascriptions. So this version is *also* exposed to the objection that threatens the first two versions, namely how could *unfree alternative* decisions and actions, which the agent is *not morally responsible* for, possibly make it the case that the agent decides and acts *freely* in the *actual* sequence and is *morally responsible* for what she actually does. However, instead of pursuing this objection further against the third, backward-looking version, Fischer prefers to prevent the resistance movement from getting off the ground. He does so by showing how to construct Frankfurt cases in which the only alternatives are, by anybody's lights, plainly not robust enough to ground moral responsibility. So imagine that Black's intervention is triggered by a sign that is clearly not action-like and prior to any action-like event that the agent could perform. This sign is something that *happens* to the agent and is obviously beyond his control, but if the agent were to show it, she would not even be able to try or begin to decide otherwise or to form a different preference. Here is the example in Fischer's own words:

> Suppose we again consider the version of the Jones and Black case in which Black can be alerted to Jones's future inclination [not to kill Smith] by the presence of some involuntary sign, such as a blush or twitch or even a complex neurophysiological pattern. So if Jones were (say) to blush red, then Black could intervene prior to Jones's doing *anything* freely and ensure that Jones indeed [kills Smith]. Here the "triggering event" (i.e., what would trigger the intervention of Black) is *not* any sort of initiating action, and thus cannot be said to be freely done. Again…this sort of triggering event appears to be not sufficiently robust to ground responsibility ascriptions.
>
> (Fischer 1994:144)

This way of constructing the example clearly seems to put to rest any "flicker of freedom" attempt to rescue the necessity of alternatives for moral responsibility, at least if this condition is read in a reasonable way, as grounding, at least partly, moral responsibility and ascriptions thereof. If, in a case like this, the agent is morally responsible for what she actually does, it is extremely implausible to hold that the possibility that she blushes or twitches is what, at least partly, grounds her moral responsibility. Thus, freedom to do otherwise (or, in Fischer's terms, regulative control) is not really required for moral responsibility. Alternatives may be present, but, to borrow a Wittgensteinian image, they turn idly, unconnected to the mechanism that gives rise to moral responsibility ascriptions.

The core of the robustness objection to the "flicker" strategy has to do with the question of the agent's *control* over the alternatives. Alternative possibilities over which the agent does not have control are not robust enough to ground her moral responsibility for what she actually does. As Fischer (2003:244) has recently put this objection against Rowe's version of the "flicker" strategy, what matters concerning robustness is not the dimension of size, but of voluntariness. The relevant control is, then, understood as voluntary. The mere fact that, in Frankfurt cases, something different happens in the alternative sequence and in the actual sequence is not enough to save PAP, unless the agent has voluntary control over her access to that alternative sequence. Some thinkers have considered this control requirement as an insuperable obstacle to the "flicker" strategy. So, according to McKenna and Widerker, in cases such as Fischer's…

> [The] alternatives present…are *not* significant for the purpose of assessing an agent's moral responsibility for her conduct in the actual world. Why? Because such alternatives are not within the scope of the *agent's control*. The prior sign…is not within the voluntary control of the agent…[A] *robust alternative*…must satisfy two conditions. One is that it is morally significant *simpliciter*. It would tell us something about the moral quality of the agent's conduct were she to have acted on it. Another is that the alternative has to be within the control of the agent. Frankfurt examples making use of prior signs thwart the flicker defense by undercutting the second condition. For this reason, we do not believe that the flicker defense is fruitful.
>
> (McKenna and Widerker 2003:8)

Leaving aside momentarily whether Widerker and McKenna are right in their rejection of the flicker strategy, the necessity of alternative possibilities for moral responsibility can be defended in a different and powerful way, which has come to be known as the Kane—Widerker dilemma, or just as the dilemma defence. Let us go into it.

The dilemma defence

Though Robert Kane had already advanced a version of the dilemma defence of PAP against Frankfurt-inspired attacks in his 1985 book *Free Will and Values,* it was David Widerker's new and forceful version (Widerker 1995) that took the dilemma defence to the forefront of the discussion.[4] Common to both versions was the insistence on decisions or choices, rather than actions, as the proper *loci* of moral responsibility for libertarians. Unlike overt actions, which are complex processes, Widerker takes decisions to be simple mental acts. His main contention is that Frankfurt-type counterexamples do not threaten PAP as applied to decisions.

As a preliminary point, remember that, according to Frankfurt, the apparent plausibility of PAP may derive from the wrong opinion that, if coercion precludes an agent's moral responsibility, it is by depriving her of alternative possibilities. It is true that, in cases of coercion, the agent cannot do otherwise, but it is also true that the agent acts as she does because of the coercion. The same causal factor, namely coercion, which excludes alternative possibilities, also explains the action. So the agent could legitimately excuse herself by claiming that she acted as she did because, given the coercion, she had no alternatives. But it is this causal history, and not the lack of alternative possibilities as

such, that rules out the agent's moral responsibility. This can be shown by reflecting on cases in which both coercion and alternative possibilities are absent, but where the circumstances that leave the agent no alternative do not cause nor causally explain what she does. Rather, the agent decides and acts for reasons of her own. In these cases, Frankfurt contends, the agent is fully morally responsible for her action. Frankfurt's counterexample to PAP is supposedly of this sort. Widerker insists on this crucial feature of Frankfurt-type examples, which he calls the IRR thesis: "There may be circumstances in which a person performs some action which, although they make it impossible for him to avoid performing that action, they in no way bring it about that he performs it" (Widerker 1995:248).

Widerker contends that no known Frankfurt-type cases satisfy IRR. He argues for this point on the basis of a Frankfurt-type example similar to that construed by Fischer against the "flicker" strategy, where the sign that triggers Black's intervention is purely involuntary and beyond Jones's control. In Widerker's example, the sign that triggers Black's intervention is Jones's *not* blushing at t1. Black knows that, if Jones blushes at t1, Jones will decide at t2 to kill Smith and that, if Jones does not blush at t1, Jones will not decide at t2 to kill Smith. Of course, in the actual sequence, Jones blushes at t1 and decides at t2 to kill Smith on his own, and Black remains inactive. Apparently, Jones lacks any robust alternatives: blushing or not blushing is plainly not robust enough to ground his moral responsibility for deciding to kill Smith in the actual sequence. Moreover, since the factor that prevents Jones from deciding otherwise, namely Black's intervention, remains purely counterfactual and in no way causes Jones's actual decision, the example seems to satisfy IRR.

So Jones is morally responsible for deciding to kill Smith, though he could not avoid that decision. And the only existing alternatives, namely Jones's blushing or not blushing, are plainly irrelevant to his moral responsibility for that decision. It seems, then, that this example, similar to Fischer's, shows PAP to be false, even for decisions. But this is only an appearance, Widerker contends. To see this, think of the sign that Black employs, in the actual sequence, as his clue for *not* intervening, namely Jones's blushing at t1. Concerning this sign, the following dilemma arises (cf. Widerker 1995:251–6). Either Jones's blushing at t1 is (or indicates a condition that is) causally sufficient for his decision to kill Smith at t2 or it is not. If it is, then Jones has no alternatives to that decision, but the circumstances that make it impossible that he decides otherwise are not purely counterfactual. Rather, they *cause* Jones's decision. So this is not an IRR situation. The decision is causally determined, and incompatibilists will plausibly refuse to accept that Jones is morally responsible for it. If, on the other hand, the sign is not causally sufficient for Jones's decision at t2, then it is hard to see how the decision is unavoidable. In this case, Jones may be morally responsible for his decision to kill Smith, but he had robust alternatives to it: he could have decided otherwise. In either case, PAP, as applied to decisions, has not been falsified.

Widerker's dilemma is a powerful incompatibilist reply to Frankfurt-type attacks on PAP. If Frankfurt-type examples are not to be contentious, it is important that the factors in them that feature in the actual causal history of the agent's decision are not, at the same time, causally responsible for her lack of alternative possibilities. As we saw, the incompatibility between determinism and moral responsibility is the conclusion of two premises, namely that alternative possibilities are required for moral responsibility and

that determinism excludes alternative possibilities. Frankfurt's aim is to show that the first premise is false. But doing this in a way that assumes the falsity of the conclusion seems question-begging. Incompatibilists need not accept that Jones is morally responsible if there is, in the actual sequence, a factor, beyond the agent's control, that is causally sufficient for the agent's decision. But, if there is no such factor, why should they accept that the decision is unavoidable? Black may intervene after Jones's decision not to kill Smith and force him to do the killing. But then it is too late to deprive Jones of robust alternatives. Moreover, concerning Jones's action, it is pertinent to resort to the "flicker" strategy and to hold, following Naylor or Davidson, that he has alternatives of some sort, for, after Black's intervention, and unlike what happens in the actual sequence, Jones does not kill Smith intentionally or on his own.

It is important to note that, in his original paper, Frankfurt was aware that assuming the truth of determinism while ascribing moral responsibility to Jones would be question-begging against incompatibilists. Black's prediction of what Jones will decide and do should not be based on the assumption of determinism. However, he would not think that his example begs that question, even if it is construed in a way similar to Fischer's or Widerker's, for in a footnote he writes that "the assumption that Black can predict what Jones will decide to do does not beg the question of determinism" (Frankfurt 1969:6, fn. 3). He asks us to suppose that Jones has faced several times the choice (between A and B) he now confronts and that he has always twitched shortly before choosing A and never before choosing B. Knowing all this, and seeing the twitch, Black could predict that Jones will be choosing A shortly afterwards. According to Frankfurt, however, we can grant all this without assuming determinism:

> This does, to be sure, suppose that there is some sort of causal relation between Jones's state at the time of the twitch and his subsequent states. But any plausible view of decision or of action will allow that reaching a decision and performing an action both involve earlier and later phases, with causal relations between them, and such that the earlier phases are not themselves part of the decision or of the action. *The example does not require that these earlier phases be deterministically related to still earlier events.*
>
> (Frankfurt 1969:6, fn. 3, my emphasis)

In this version of the example, the alternatives of Jones's twitching or not twitching are relevantly similar to Jones's blushing or not blushing in Widerker's example. And Frankfurt does not seem to think that the example begs the question of determinism by allowing Jones's twitching at a certain time in the process of deliberation to be causally sufficient for his decision (to kill Smith).[5] The reason seems to be that by "determinism" he understands the general thesis that a complete description of the state of the world at any instant in the past, together with the laws of nature, implies all truths. In this sense, letting the twitch be causally sufficient for Jones's decision does not assume determinism provided that the twitch itself is not inevitable given the past and the natural laws. Whether Jones twitches or not may be genuinely undetermined, although, once he twitches, his decision to kill Smith is inevitable.

However, another version of the dilemma may be stated on the basis of the general thesis of determinism. We can find this version in a recent book on free will by Laura Waddell Ekstrom, an incompatibilist with regard to determinism and moral responsibility

(Ekstrom 2000). According to Ekstrom, in Frankfurt's example, incompatibilists should not accept the judgement that Jones is morally responsible for what he does. Rather, the right reaction is to remain agnostic about Jones's moral responsibility, for they are not given all the relevant information. In particular they are not told whether determinism is (assumed to be) true. The disjunction that starts the dilemma is: either determinism is true or it is not. And Ekstrom writes:

> If it is true, then Jones cannot do otherwise than kill Smith, but Jones is not morally responsible for killing Smith… Alternatively, suppose that determinism is not true. Then, Jones could not have done otherwise than kill Smith, by hypothesis, yet there may be indeterminism in an appropriate place or places to ground Jones's moral responsibility for the act of killing…[H]e could have decided otherwise, or could have tried to decide otherwise, or could have formed a different preference.
>
> (Ekstrom 2000:197)

Not knowing whether determinism is assumed to be true or not, we do not have enough information about Jones's circumstances to give a clear verdict about his moral responsibility. This version of the dilemma defence of PAP is importantly different from Widerker's version. In Chapter 1, we argued for the truth of the second premise of the Incompatibilist Argument, namely that determinism rules out alternative possibilities. Suppose, as seems plausible, that this premise is true and that determinism is assumed to hold in Frankfurt's example. This would rule out *any* alternatives, including such alternatives as Jones's blushing or not blushing, and his twitching or not twitching. But then Black's lurking presence would become superfluous: the actual deterministic chain of events would be enough to ensure Jones's decision. Incompatibilists would be fully justified in rejecting Jones's moral responsibility. But, as we have seen, Frankfurt seems to think that assuming the general thesis of determinism would be question-begging and there are reasons to think that he does not make such an assumption.[6] In addition to his remark that his example does not require that earlier phases of Jones's deliberation "be deterministically related to still earlier events", he also admits that "it is in a way up to him [Jones] whether he acts on his own or as a result of Black's intervention. That depends upon which action he himself is inclined to perform", though in either case "he performs the same action" (Frankfurt 1969:8). It clearly seems, then, that Frankfurt's example does not assume the truth of determinism.

So what could Frankfurt respond to Ekstrom's dilemma? He would embrace the second horn, and could contend that in an indeterministic world there can be causal deterministic sequences of events. The relationship between Jones's sign that he is going to decide to kill Smith (his twitch) and his actual decision could be such a sequence. And he could also insist, following Fischer, that the alternatives that are left are not robust enough to ground Jones's moral responsibility.

This response to Ekstrom's dilemma, however, would lead Frankfurt to confront Widerker's. Widerker would contend that assuming the existence of a causal deterministic sequence between Jones's twitching (or blushing, in Widerker's example) and his decision would *also* beg the question against the incompatibilist. Even if the

twitch itself is not causally determined (it might not have taken place) and the general thesis of determinism is not assumed, Jones's decision is still causally determined, and for an incompatibilist this is a reason for denying that the agent is morally responsible for it.

At this point, however, Frankfurt's position can, in my view, be defended as follows. Rejecting *any* causally sufficient condition of an agent's decision as incompatible with her moral responsibility for this decision may also, from the compatibilist side, appear question-begging and unmotivated. To see this, it is worth comparing Fischer's and Frankfurt's examples. Although, in both cases, the sign that tells Black about Jones's future decision is involuntary, Fischer assumes that this sign takes place before Jones forms any inclination or preference towards that particular decision. If a sign of this sort, which the agent has no control over, is causally sufficient for his later decision, then, even if the sign itself is causally undetermined, it is certainly reasonable for an incompatibilist to reject the agent's moral responsibility, for that sort of causal determination deprives the agent of any effective control over the decision she makes. But even a compatibilist might concur with the incompatibilist in this judgement, for in such a context the agent may lack reflective control over her decision, in that it derives from a process that contains "brute" causal links between merely physiological, contentless states, such as blushes, and intentional states and acts, such as reasons, preferences and decisions.

Frankfurt's example, however, looks importantly different. In it, the sign (a twitch) takes place when the agent was "about to decide", at a time, then, very close to the decision and after the agent has gone through a good deal of deliberation. Here it is plausible to think of the twitch as an external, epiphenomenal manifestation of a certain phase of the agent's deliberation process. Suppose that this phase corresponds to the agent's practical judgement that, all things considered, killing Smith is better than the alternative.[7] And suppose now that this judgement, though causally undetermined, is, once the agent forms it, causally sufficient, *ceteris paribus,* for her decision to kill Smith. In these circumstances, it would be unreasonable, for an incompatibilist, to reject the agent's moral responsibility for her decision. After all, the agent had (regulative) control over her practical judgement: she ought to and could have formed a different practical judgement, and this looks like a robust alternative.

We can conclude that the mere presence, in the actual sequence of a Frankfurt-type example, of a causally sufficient factor of the agent's decision does not, for an incompatibilist, justify a rejection of the agent's moral responsibility for that decision. Whether the rejection is justified depends on which factor it is and on whether the agent has appropriate control over its presence or absence. In Fischer's example the agent lacks this control, but in our reconstruction of Frankfurt's example she does not. It is not accidental that Widerker formulates his dilemma on the basis of a Frankfurt example similar to Fischer's, where the alternatives (blushing or not blushing) are clearly beyond the agent's control.

The dilemma defence of PAP puts important pressure on the construction of Frankfurt cases. To conclude this section, let us look at some restrictions on plausible Frankfurt cases that derive from our discussion about the dilemma defence of PAP.

First, the general thesis of determinism should not be assumed to be true in the actual sequence of a Frankfurt case, for if it is, as we have seen, incompatibilists will, reasonably enough, reject the agent's moral responsibility. Even someone like Derk Pereboom, who thinks that alternative possibilities are not required for moral responsibility, holds that, none the less, determinism rules out moral responsibility, for it rules out the sort of control or self-determination that would be needed for moral responsibility. Second, in the actual sequence, the example may contain a factor that is causally sufficient for the agent's decision, but only if the agent has appropriate control over such a factor. This excludes as illegitimate merely physiological, contentless states, such as blushes or twitches, unless they are purely epiphenomenal manifestations of causally sufficient factors over which the agent has appropriate control, such as practical judgements. An example that flouts this restriction will inevitably fall prey to Widerker's dilemma. But even compatibilists may justifiably question the agent's moral responsibility, on the basis that she lacks appropriate reflective control over her deliberation and decision. Bear in mind that an important source of the intuition that the agent is responsible in Frankfurt-type cases is the fact that the agent reaches her decision on the basis of her own reasons and through a normal process of deliberation that involves justification relationships between reasons, decision and action.[8]

It is essential to a Frankfurt case that the agent's decision and action are unavoidable, that robust alternative possibilities are absent, and that the agent is clearly morally responsible for what she decides and does. The challenge for Frankfurt theorists is to construct cases with these features while, at the same time, respecting the aforementioned restrictions. In the next section we shall examine a recent attempt to meet this challenge.

Actual blockage cases

An important proposal to deal with the problems raised by the dilemma defence of PAP is to replace the counterfactual factors featured in classical Frankfurt cases by actual blocking mechanisms that in fact prevent any alternatives from arising *without, however, causing them not to arise,* so that the situation is supposed to satisfy IRR. In these examples the agent shows *no prior sign* of her later decision, and the disjunction that starts Widerker's dilemma loses its footing. Moreover, the agent is supposed to reach her decision through an indeterministic process: the general thesis of determinism is not assumed to be true. However, owing to the blocking device, what the agent decides and does is the only thing she can actually decide and do.

Cases of this sort have been proposed by David Hunt (Hunt 2000) and by Alfred Mele and David Robb (Mele and Robb 1998).

Hunt thinks it is worth exploring whether "the unavoidability essential to a Frankfurt scenario" necessarily has to rest "on a counterfactual device" (Hunt 2000:217). Hunt suggests it does not. In fact, he thinks that Locke's classical example of a man who gladly remains in a room which he cannot actually get out of provides support for this answer. Of course, Locke's example as it stands cannot yield the wanted results, for it still leaves many alternatives open which are relevant to the agent's moral responsibility. Hunt then conceives of some other ways in which "unavoidability does not wait upon a counterfactual trigger and so can extend to all the agent's actions, leaving no alternate possibilities to ground moral responsibility" (Hunt 2000:217). Of these ways, by far the

most promising against expected rejoinders by alternative possibility defenders is to show how to construct blockage cases. The general structure of a blockage case is nicely shown by Hunt himself in the following text:

Imagine then a mechanism that blocks neural pathways rather than doorways. Suppose that the actual series of Jones's mental states leading up to the murder of Smith is compatible with PAP, except that the mechanism is in operation. The mechanism is not intervening directly in the series itself; it is allowing the series to unfold on its own, but simply blocking all alternatives to the series. Of course it can't block alternatives *in response* to the way the series is unfolding, because then the blockage would be coming too late to have any effect on the avoidability or unavoidability of Jones's actions. Instead, the mechanism blocks alternatives in advance, but owing to a fantastic coincidence the pathways it blocks just happen to be all the ones that will be unactualized in any case, while the single pathway that remains unblocked is precisely the route the man's thoughts would be following anyway (if all neutral pathways were unblocked). Under these conditions, the man appears to remain responsible for his thoughts and actions…

(Hunt 2000:218)

A fantastic coincidence, indeed. But we are supposed to play with conceptual possibilities, and this would seem to be one. Let us examine the case by using a concrete example of this rather general structure. The example has been designed by Mele and Robb:

At *t1,* Black initiates a certain deterministic process *P* in Bob's brain with the intention of thereby causing Bob to decide at *t2* (an hour later, say) to steal Ann's car. The process, which is screened off from Bob's consciousness, will deterministically culminate in Bob's deciding at *t2* to steal Ann's car unless he decides on his own at *t2* to steal it… The process is in no way sensitive to any "sign" of what Bob will decide. As it happens, at *t2* Bob decides on his own to steal the car, on the basis of his own indeterministic deliberation about whether to steal it, and his decision has no deterministic cause. But if he had not just then decided on his own to steal it, *P* would have deterministically issued, at *t2,* in his deciding to steal it. Rest assured that *P* in no way influences the indeterministic decision-making process that actually issues in Bob's decision.

(Mele and Robb 1998:101–2)

The conclusion seems to be, again, that Bob is responsible for stealing Ann's car. One may wonder what would happen if, at *t2,* the decision arrived at by Bob's indeterministic decision-making process, call it *"x",* is the decision *not* to steal Ann's car. Mele and Robb address this issue. Their response, in general terms, is as follows. In case of conflict, *P,* the deterministic process initiated by Black in Bob's brain, would prevail. But, if there is no conflict, the decision to steal Ann's car will be caused by *x,* Bob's indeterministic decision-making process, which will then prevail over *P*. But this looks a bit too nicely arranged, and again one may wonder why in the first situation *P* prevails over *x* while in the second it is the other way around. Mele and Robb provide an answer in the form of a story, where *N2* is a "decision node" in Bob's brain whose "lighting",

on being "hit" by *P* or *x,* represents his decision not to steal Ann's car at *t2* (the lighting of *N1* would represent his decision to steal the car). According to this story, by *t2 P* has blocked *N2* without affecting the unfolding of process *x*. So, if *x* were to hit *N2* at *t2, N2* would not light up. More exactly, by *t2 P* has blocked all decision nodes in Bob's brain that are contrary to a decision at *t2* to steal Ann's car. Again, of course, this blockage does not cause Bob's decision at all, since at *t2* his own indeterministic decision process hits *N1* and he decides on his own to steal Ann's car.

Now, are we forced to accept the intended conclusion, namely that Bob is morally responsible for deciding to steal Ann's car even if he had no alternatives to that decision? I do not think that our intuitions here speak up as clearly for this conclusion as in classical Frankfurt cases. Even if we accept that *P* does not cause Bob's decision, *P* is *actually,* not merely counterfactually, blocking any alternative to that decision. How shall we exclude the possibility that an actual, deterministic blocking mechanism in Bob's brain is not, in any way at all, influencing Bob's decision-making process? Stipulating that it is not does not seem to be enough in this case to dispel our doubts.

But Mele and Robb have another way of presenting his case "from an intuitively appealing perspective", a perspective that, in a note, they acknowledge was recommended to them by John Martin Fischer. Here it is:

> Subtract Black and *P* from our scenario and imagine that what happens at Bob's indeterministic world is that *x,* Bob's indeterministic decision-making process, indeterministically issues at *t2*—in some way favored by libertarians—in his decision to steal the car. Plainly, there is no deterministic cause of Bob's decision in this case. Now add Black and *P* to the scenario in just the way we have done. At *t2,* process *x* issues in the same indeterministic way in Bob's decision: by hypothesis, Black and *P* do not influence *x*. Although at *t2* Bob cannot do otherwise than decide to steal the car, nothing warrants the claim that his decision is deterministically caused.
>
> (Mele and Robb 1998:108)

So the idea seems to be this. We are certainly willing to accept that Bob is morally responsible for his decision in the first scenario depicted above, where no blocking mechanism is operating. Our intuitions are quite clear in this case. Now, suppose that we refuse, or at least are reluctant, to accept that Bob is morally responsible for his decision to steal Ann's car in the original example, when the blocking mechanism is in operation. This is the second scenario. Then the challenge is for us to find a *relevant* difference between these two scenarios to justify the difference in our respective judgements about Bob's moral responsibility. The difference has to be relevant, for difference there is: a blocking mechanism in the second case that is absent in the first. But this difference, Mele and Robb clearly think, is not relevant to justify the difference in our respective judgements, for the actual causal story is the same in both cases.

Mele and Robb claim that their example meets objections that may affect classical Frankfurt cases: there is no sign showing which would introduce some alternative possibilities into the picture, or not; determinism is not assumed; and, as they also insist, the example "is what Widerker [1995:248] calls an 'IRR-situation': there are 'circumstances in which' Bob decides to steal Ann's car that 'make it impossible for him to avoid' deciding to do this but 'in no way bring it about that' he decides to do this"

(Mele and Robb 1998:108). So, we may add, what should prevent one from accepting Bob's moral responsibility except perhaps arcane libertarian prejudices?

Mele and Robb's example, however, is not free of problems. Concerning Widerker's version of the dilemma defence, bear in mind that *P,* the deterministic causal process initiated by Black in Bob's brain, has to remain active until the very moment of Bob's decision, so that both processes, *P* and *x,* hit decision node *N1* at *t2,* for, if *P* were deactivated before *t2,* Bob could decide not to steal Ann's car. And, as Widerker writes: "This being the case, it is hard to see how *x* could prevent *P* from deterministically causing Bob's decision. It would be simply too late for *x* to accomplish that" (Widerker 2003:56). It seems, then, that the situation depicted is not an IRR situation, so that the example falls prey to Widerker's dilemma.

In order to respond to this objection, Mele and Robb (2003:134–6) have to postulate a special kind of pre-emption, which allows the indeterministic process *x* to pre-empt the deterministic process *P* even when the two processes hit decision node *N1* simultaneously. They acknowledge that this kind of pre-emption is "unusual and certainly not…standard" (Mele and Robb 2003:134). But if this sort of pre-emption were to appear as not possible, that would undermine the example.

But Mele and Robb's example also faces problems concerning Ekstrom's version of the dilemma defence: as Derk Pereboom (Pereboom 2000, 2001) has argued, Mele and Robb might be assuming determinism. Pereboom constructs the following two-scenarios case, which involves an atom. Imagine a universe correctly described by Epicurean physics, in which all that ultimately exists is atoms and frictionless void. As is known, according to this physics atoms do not have completely deterministic trajectories: from time to time they suffer uncaused swerves. Suppose they naturally fall downwards. Now, this is the first scenario: "*Situation C*. A spherical atom is falling downward through space, with a certain velocity and acceleration. Its actual causal history is indeterministic because at any time the atom can be subject to an uncaused swerve. Suppose that the atom can swerve in any direction other than upwards. In actual fact, from *t1* to *t2* it does not swerve" (Pereboom 2001:17). The second scenario is as follows: "*Situation D*. The case is identical to *C,* except that the atom is falling downward through a straight and vertically oriented tube whose interior surface is made of frictionless material, and whose interior is precisely wide enough to accommodate the atom. The atom would not have swerved during this time, and the trajectory, velocity, and acceleration of the atom from *t1* to *t2* are precisely what they are in *C*" (Pereboom 2001:17). Pereboom comments on this case as follows:

> One might initially have the intuition that the causal history of the atom from *t1* to *t2* in these two situations is in essence the same. However, this intuition could be challenged by the fact that the restrictions present in *D* but not in *C* may change this causal history from one that is essentially indeterministic to one that is essentially deterministic. For since the tube prevents any alternative motion, it would seem that it precludes any indeterminism in the atom's causal history from *t1* to *t2*. And if the tube precludes indeterminism in this causal history, it would appear to make the causal history deterministic.
>
> (Pereboom 2001:18)

According to Pereboom, his line of argument does not show in a conclusive way that blockage cases do actually assume determinism in the actual sequence, but it certainly suggests that they might be doing this. Kane (2000) and Ekstrom (2002) also suspect that, in Mele and Robb's example, Bob's decision is deterministically caused. In a recent paper, Mele and Robb (2003:130) face these criticisms. They acknowledge that *P,* together with the laws of nature and the circumstances at *t2,* entails Bob's decision at *t2* to steal Ann's car. So Pereboom, Kane and Ekstrom might argue that this shows Bob's decision to be deterministically caused by *P*. Mele and Robb accept that this conclusion "would follow immediately on a nomic subsumption model of causality" (2003:130), as well as on certain counterfactual models. Their response is to reject such models of causality. But this makes the success of their example depend on this contentious issue about the metaphysics of causality.

As a result of the discussion so far, the suspicion arises that blockage cases may be implicitly violating some restrictions, related to causality and determinism, which, as we pointed out in the preceding section, non-contentious Frankfurt cases should respect. However, I think that a much stronger case against the blockage strategy can be made on the basis of the requirement of reflective self-control and responsiveness to reasons. This condition can be made to weigh heavily, in fact decisively, against "blockage" attempts to reject the necessity of alternative possibilities for moral responsibility. I shall focus on Mele and Robb's example, but the argument applies to other versions, such as Hunt's.

First of all, though this is not an essential point, it might seem that the way in which Mele and Robb depict decisions in their story is a bit too simple. They talk about neural "decision nodes". The blockage affects all decision nodes incompatible with Bob's decision at *t2* to steal Ann's car. Decisions, however, and especially decisions to which moral responsibility ascriptions paradigmatically apply, are preceded by and based on consideration of reasons. But in Mele and Robb's story, while some decision nodes are blocked, nothing is said about *P's* blocking of the "reason pathways", as we could call them, that speak for those decisions. Now imagine the following case. Close to *t2,* when decision nodes incompatible with Bob's decision at *t2* to steal Ann's car have been already blocked by *P* (remember the story) and Bob's decision to steal Ann's car is therefore unavoidable, Bob is told by a friend of his that he has put a bomb in Ann's car and that it will explode if someone tries to get into the car. (One may substitute for this any decisive, overwhelming reason one can think of for Bob not to steal Ann's car.) What happens then to Bob? He clearly sees that he is going to die if he decides to steal Ann's car; he clearly sees that he has a decisive reason to decide not to steal it; but when he tries to decide not to steal the car, he finds himself unable to do it. Poor Bob. He sees how his reasons for decision and his decisions come dramatically apart.

So this is our story. What it shows is that the blockage is not without consequences for Bob's decision-making capacity. It affects his dispositions to decide according to those reasons he finds better. Fischer and Ravizza (1998:41–6) have proposed one way, which they call "weak reasons-responsiveness", of specifying rationality-related conditions of moral responsibility. But whether or not this way is fully satisfactory, there is little doubt that something like this must be a necessary condition of moral responsibility. Fischer and Ravizza's idea is roughly as follows: weak reasons-responsiveness holds just in case, keeping the agent's actual mechanism of deliberation and decision-making constant, there are some possible scenarios or possible worlds in which there is a sufficient reason

to decide and do otherwise, the agent recognizes this reason and decides and does otherwise. Think of someone who decides to steal a book and does so (the example is Fischer's). Fischer writes: "If (given the operation of the actual kind of mechanism) he would persist in stealing the book even if he knew that by so acting he would cause himself and his family to be killed, then the actual mechanism would seem to be inconsistent with holding him morally responsible for his action" (Fischer 1994:167). But this seems to be precisely the case with Bob. And this undermines the judgement that, when the blockage is in operation, he is morally responsible for his decision.

Someone might try to resist this objection by claiming that there are some possible worlds in which Bob recognizes the reason for not stealing Ann's car and decides and acts accordingly. These worlds are those in which there is no blockage in operation. The response is that these worlds are not such that the actual mechanism operates in them. The actual mechanism with which Bob reaches his decision to steal Ann's car has some pathways blocked. So removing the blockage changes the actual mechanism. Note that this is not the case in classical Frankfurt cases. In these cases, there are worlds, namely those in which the counterfactual intervener is not present, in which the actual mechanism operates and the agent is appropriately responsive to reasons. This is precisely the difference between ruling out alternatives by means of merely counterfactual factors and ruling them out through actual blocking devices.

Our example shows that what matters for moral responsibility is not only the actual causal history of a decision or action, but also the truth or falsity of certain counterfactuals about how the agent would decide and act if certain reasons were present. Even if determinism could be shown not to have been assumed and, as a matter of fact, Bob's decision were not causally determined, it could not be said to be free, in the sense relevant to moral responsibility, for it issues from an impaired, non-reasons-responsive mechanism of deliberation and decision-making. We see that, even if Bob were given an absolutely decisive reason for not deciding in a certain way, he would still decide that way, by virtue of features of the mechanism with which he makes his decision. His practical rationality is thus seriously impaired, and this, for incompatibilists *and* compatibilists alike, undermines an agent's moral responsibility. The considered judgement, then, about Bob's case is that he is not morally responsible for his decision. This may explain our reluctance, in actual blockage cases, to accept that the agent is fully morally responsible.

We said above that Mele and Robb's view of decisions looks too simple. So one may think of modifying their example so that the blockage does not only affect those "decision nodes" contrary to Bob's decision to steal Ann's car at *t2,* but also all contrary "reasons-appreciation nodes", as we may call them. Then Bob will become insensitive to any reason that goes against that decision. In this case, were he presented with our (or another) decisive reason not to steal Ann's car, he would not feel the split between that reason and his decision, for he would not even see the force of this reason. But this does not solve the problem concerning his moral responsibility; rather, it makes it worse, for the impairment of his deliberation and decision-making mechanism extends even further than in the former case. Now he is not only unable to decide as a decisive reason recommends, but he is also unable to appreciate the force of such a reason.

These considerations suggest an important point, which goes beyond the limits of the issue we are discussing. Notice that, by actually ruling out any alternatives of decision, the blockage has also affected the agent's capacity for practical reasoning. This strongly suggests that alternative possibilities and practical rationality are not independent of one another. Or, more precisely, that alternative possibilities are an essential aspect of rationality. If alternative possibilities are effectively swept away, rational capacities for deliberation and decision-making are seriously affected as well. This point will become prominent in the final chapter.

We can now answer Mele and Robb's challenge. There is a difference between the two scenarios they depict, which may explain why our judgement about the agent's moral responsibility becomes unstable, to say the least, when blockage is added to the picture. The difference is that, in this latter case, the agent's capacities and dispositions for practical deliberation are seriously impaired by the blockage. And this justifies the difference in our judgement about Bob's moral responsibility when we move from one scenario to the other.

In the light of the preceding arguments, I think it is fair to say that the actual blockage strategy against the alternative possibilities condition is unlikely to succeed. Let us insist that classical Frankfurt cases do not face the rationality problem we raised for blockage cases, for in the former, the actual mechanism of deliberation and decision is supposed to be adequately responsive to reasons. The mechanism in the counterfactual situation may not be reasons-responsive, but this mechanism is not the same as that which operates in the actual sequence, since now the counterfactual factor has taken over. We think, then, that Hunt is wrong when he writes that "the unavoidability essential to a Frankfurt scenario does not have to rest on a counterfactual device" (Hunt 2000:217).

Other attempts to deal with the dilemma defence of PAP are closer to classical Frankfurt examples. The new cases feature a prior sign and a counterfactual intervener, and, by their very structure, they contain alternatives of *some* sort. The claim will be that these alternatives, present as they may be, are not robust enough to ground the agent's moral responsibility. Let us examine one prominent attempt on these lines, developed by Derk Pereboom.

Robustness, determinism and the dilemma defence

Blockage cases have been designed as a response to the dilemma defence of PAP. They are intended to show that an agent's decision may be unavoidable even if it is not deterministically caused. As we have argued, it is not clear that they are successful on this account. If they are not, they do not actually escape some form of the dilemma defence. However, we have contended that, even if they are, the actual blockage of alternative pathways damages the reasons-responsiveness of the actual mechanism of decision-making, and this undermines the judgement that the agent is morally responsible for her decision. Though this objection has not been appreciated in the literature, I think it is decisive against blockage cases. However, classical Frankfurt cases, where the factor that supposedly rules out alternative possibilities remains purely counterfactual, seem to be immune to it. New attempts to rescue Frankfurt's criticism of PAP from the dilemma objection are intended to respect this feature. Derk Pereboom's proposal is prominent among them. As we saw, Pereboom suspects that blockage cases may be guilty of

introducing determinism in the actual history of the agent's decision and he is especially anxious to avoid this mistake. Pereboom's example features Joe, who is thinking of illegally claiming a tax deduction:

Tax Evasion, Part 1. Joe is considering whether to claim a tax deduction for the substantial local registration fee that he paid when he bought a house. He knows that claiming the deduction is illegal, that he probably won't be caught, and that if he is, he can convincingly plead ignorance. Suppose he has a very powerful but not always overriding desire to advance his self-interest… Crucially, his psychology is such that the only way that in this situation he could fail to choose to evade taxes is for moral reasons… In addition, it is causally necessary for his failing to choose to evade taxes in this situation that he attain a certain level of attentiveness to these moral reasons. He can secure this level of attentiveness voluntarily…[9]

(Pereboom 2003:193)

Now, it is an essential ingredient of the way Pereboom depicts the situation that Joe's attaining the required level of attentiveness to moral reasons, though causally necessary, is, however, *not causally sufficient* for his failing to choose to evade taxes, so that his attaining that level does not ensure this result: "If he were to attain this level of attentiveness, Joe could, with his libertarian free will, either choose to evade taxes or refrain from so choosing (without the intervener's device in place)" (Pereboom 2003:193). In fact, Joe is supposed to be a "libertarian free agent". But a counterfactual factor, which never needs to interfere in Joe's process of deliberation and decision-making, ensures that Joe will choose to evade taxes, for

…a neuroscientist now implants a device which, were it to sense the requisite level of attentiveness, would electronically stimulate his brain so that he would choose to evade taxes. In actual fact, he does not attain this level of attentiveness, and he chooses to evade taxes while the device remains idle.

(Pereboom 2003:193)

The conclusion we are supposed to draw is that Joe is morally responsible for choosing to evade taxes though he could not have chosen otherwise. Let us comment on some features of the example in order to see whether this conclusion is warranted.

First of all, if the example is going to work against the dilemma defence, the alternatives that are present, namely Joe's attaining a certain level of attentiveness to moral reasons or not, have to be genuinely undetermined, and this does indeed seem to be Pereboom's view. The example seems to work well against Ekstrom's version of the dilemma defence. The general thesis of determinism is not assumed, but, Pereboom will hold, the alternatives that are left are not, contrary to Ekstrom's contention, sufficiently robust to ground the agent's moral responsibility for his decision to evade taxes. This decision is unavoidable and ensured by the counterfactual device. It is less clear, however, that the example can also succeed against Widerker's version of the dilemma defence. The way in which Pereboom intends to escape it is to take the sign that triggers the device's activation, namely Joe's attaining a certain level of attentiveness to moral reasons, to be only causally necessary, and not causally sufficient, for his failing to

choose to evade taxes. That is, but for the presence of the counterfactual factor, he might still choose to evade taxes even if he attains the required level. Pereboom seems to think that this is enough to satisfy the condition that the alternatives should not be ruled out by assuming that there is a causally sufficient condition of the agent's choice. But Widerker could object that, if Joe's attaining the requisite level of attentiveness is causally necessary for his failing to choose to evade taxes, then the fact that he does *not* attain that level, as is actually the case in the example, together with the other actual circumstances of the situation, is causally sufficient for (causally determines) his *not* failing to choose to evade taxes, that is, his actual choice to evade them. But then, Widerker would contend, this is not an IRR situation: the circumstances that make it impossible that the agent chooses otherwise cause the choice. As a result, libertarians need not accept Joe's moral responsibility.

This looks like a powerful objection, but Pereboom may have the resources to resist it (cf. Pereboom 2003:195). Let us remove the counterfactual intervener from the scenario, since it does not have any real causal influence on Joe's decision, and ask whether the fact that Joe does not attain the relevant level of attentiveness at a certain time *t1* in his process of deliberation, prior to his choice at *t2,* together with the other circumstances of the actual situation, is causally sufficient for Joe's choice, at *t2,* to evade taxes. The answer seems to be that it is not. It will be only if it were determined that Joe will not attain the required level of attentiveness between *t1* and *t2*. But this is not determined. As Pereboom writes: "There is no relevant time at which refraining from deciding to evade taxes in the future is impossible for Joe, since he can always achieve the right level of attentiveness" (Pereboom 2003:195). That Joe attains this level or not is undetermined and open to him until the very moment of his choice. If Joe attains this level, he can still refrain from deciding to evade taxes, or decide to evade them instead (without the counterfactual intervener in place). If he does not attain this level, he will certainly choose to evade taxes, but, as we have argued, this choice, though certain, is not causally determined. Joe can attain the relevant level at any time throughout his whole process of deliberation, which keeps the alternative possibilities alive until the very instant of his choice. This choice, however, is not merely probable, for, if the only condition that could lead Joe to fail to evade taxes were to hold, the counterfactual device would be activated, and Joe would still choose to evade taxes. It seems, then, that Joe's decision is unavoidable, yet causally undetermined. If he is morally responsible for his choice, it seems that PAP, as applied to decisions, is false.

Fischer's example, in which the alternatives take place before the agent can form any preference or do anything freely, falls prey to Widerker's dilemma. In Pereboom's example, however, the agent retains the power to attain the required level of attentiveness throughout his whole process of deliberation. This feature seems crucial for its ability to escape the dilemma. Moreover, Pereboom's remark that Joe "can secure this level of attentiveness voluntarily" seems to grant the agent an appropriate degree of control over the alternatives, unlike what happens in Fischer's example. Moreover, unlike blushes or twitches, moral reasons have a content that gives them a legitimate role in practical reasoning. So they do not introduce brute causal chains in the process of deliberation and do not raise doubts about the rationality of that process which might undermine the ascription of moral responsibility to the agent. We have also seen that, unlike what

happens in blockage cases, the reasons-responsiveness of the actual mechanism of decision-making is not impaired, given the counterfactual character of the intervening factor.[10]

If, as seems plausible from the preceding considerations, Pereboom's example can resist the dilemma defence, and the agent is morally responsible for his choice, even if it is unavoidable, the only option for a libertarian in order to rescue the alternative possibilities condition would seem to be to embrace some version of the "flicker" strategy, pointing to the alternative that is open to Joe, namely his attaining a certain level of attentiveness to moral reasons, and to the corresponding alternative sequence triggered by Joe's attaining that level.

Against this libertarian move, Pereboom's contention is essentially Fischer's, namely that the alternatives that are left to the agent are not robust enough to save PAP: they are not explanatorily relevant to our ascription of moral responsibility to the agent for his actual choice. As Pereboom points out, "the deeper intuition underlying the alternative-possibilities requirement is that if, for example, an agent is to be blameworthy for an action, it is crucial that he could have done something to avoid this blameworthiness" (Pereboom 2001:19). This is perhaps a convoluted way of expressing the intuition behind the alternative possibilities condition, but it seems true that, if this condition is widely taken to hold, it is not only by virtue of an aesthetic preference for forked paths. As we have insisted, this condition is closely related to the control we want to have over the moral responsibility we bear for what we decide and do. In connection with this, it also plays an explanatory-cum-justificatory role in moral responsibility ascriptions. And this is what underlies the robustness objection: even if alternatives are present in Frankfurt cases, they must be explanatorily relevant to ascriptions of moral responsibility if one wants to defend PAP by appealing to them. Let us see how Pereboom spells out this notion of robustness or explanatory relevance.

Pereboom arrives at his favoured definition of robustness on the basis of some recent proposals to resist Frankfurt-type attacks on PAP, put forward by Michael Otsuka (1998), Keith Wyma (1997) and Michael McKenna (1997). These proposals are new versions of the "flicker" strategy. We shall take Wyma's contention as representative of them all. He holds (Wyma 1997) that a person is morally responsible for something she has done, A, only if she has not done something she could have done, B, such that doing it would have made her not morally responsible for A. In standard Frankfurt cases, Wyma claims, this condition is not falsified, for, though the agent is morally responsible for what she did, there is something she could have done that would have triggered the device's intervention and would have rendered her not responsible for what she did. Pereboom contends that, even if conditions of this sort, suitably refined, might be shown to be necessary for moral responsibility and free of counterexamples, they still fail to respect the explanatory relevance criterion. Think, in effect, that Wyma's condition might be trivially satisfied in cases in which the alternative possibility not chosen by the agent is clearly irrelevant, even by the agent's own lights, to her moral responsibility for what she did. For example, think of Joe, the tax evader. Suppose that he could fill in the tax deduction form with either of two pens but, unbeknown to him, one of the pens contains a bomb that would explode and seriously injure him if he used it, so that the deadline for claiming a deduction would be over when he recovered. He actually chooses the other, fills in the form and applies for the tax deduction. He is morally responsible for this act,

and there is an alternative that, had he chosen it, would have made him not responsible for this act, simply because it would not have been performed. But this alternative is clearly insufficient to ground his moral responsibility. It has no role to play in explaining his moral responsibility for evading taxes. One way of expressing this is to say that choosing that alternative would exempt him from responsibility for evading taxes just by luck. This shows that an epistemic condition should also be included in an appropriate criterion of robustness: not only has the alternative to be such that, were the agent to choose it, she would thereby be exempted from moral responsibility, but she also has to *understand* that this would be the case. Finally, since actions are not always in the agent's hands, Pereboom takes robust alternatives to be willings. This is Pereboom's final considered notion of robustness:

Robustness. For an alternative possibility to be relevant to explaining why an agent is morally responsible for an action, it must satisfy the following characterization: she could have willed something different from what she actually willed such that she understood that by willing it she would thereby be precluded from moral responsibility for the action.
(Pereboom 2001:26; cf. Pereboom 2003:194)

Now, if in the fantasy about the bomb-pen the alternative clearly does not satisfy this criterion, neither does the alternative in the original tax evasion case, according to Pereboom. Joe could have willed to attain a certain level of attentiveness to moral reasons not to evade taxes. Had he succeeded, the device would have been activated, and he would not have been responsible for evading taxes. But this would have happened by luck, for Joe, being ignorant of this fact, did not understand that by taking this alternative path he would be exempted from responsibility.

Although Linda Zagzebski is a libertarian, she endorses essentially Pereboom's position as well. But the way she makes the case throws more light on the robustness objection. She refers to Pereboom's original example (cf. Pereboom 2001:19), in which the alternative is conceived, roughly, as Joe's thinking of a moral reason, rather than as his attaining a level of attentiveness to such a reason. She writes:

If the agent could have entertained the thought voluntarily, then it is true that there is something she could have done that would have precipitated the action of the machine, thereby rendering her blameless, but her blamelessness would have little to do with what she did. Surely, when the machine does not need to intervene and in ordinary situations in which there is no machine in existence we do not blame her because she could have had such a thought. She is to blame because her act was the result of a libertarian free choice, not because she might have had a moral thought which would have been necessary but not sufficient for deciding not to commit the act. Having a moral thought is not exercising the kind of power relevant to her blameworthiness…
(Zagzebski 2000a:242)

Viewing things from the perspective of blamelessness, think of the circumstances we can plausibly allege to exempt an agent from responsibility for doing A. Provided that she has done A, it is clear that we cannot preclude her from responsibility for this action by saying that she thought of (or was attentive to) a moral reason for not doing A. And this

also holds when it is the agent himself who tries to get rid of responsibility. Think of Joe, the tax evader. If, in a normal situation where no device is lurking, he finally decides to evade taxes and does so, he would not succeed in exempting himself from responsibility if he claimed that he had thought of or was attentive to a moral reason for not evading taxes. This is simply not the right claim to avoid responsibility. Zagzebski has rightly pointed to moral luck in the context of this discussion (cf. Zagzebski 2000a:245). In fact, maybe contrary to what we initially tend to think, agents involved in Frankfurt cases are very lucky from a moral point of view. They can get rid of responsibility very easily. Joe, in Pereboom's example, could have evaded both taxes *and* responsibility if he had paid attention to a moral reason against his action, for this would have activated the device. No agent in normal situations could discharge her moral duties with so little effort. But this strongly suggests that this alternative, present as it may be, is not robust and cannot be what grounds—even partially—Joe's moral responsibility for choosing to evade taxes.

So Pereboom's example seems to succeed against the "flicker" and the dilemma defence of PAP, and to reinstate the essential core of Frankfurt's position. In the next section, however, we shall try to show that, in spite of recent criticisms, even by libertarians, the "flicker" defence can succeed against the robustness objection and that PAP can be effectively vindicated on that basis.

Rescuing the "flicker" strategy

The dilemma defence of PAP succeeds in discarding Frankfurt cases in which the alternatives are mere happenings, beyond the agent's control. But Pereboom's example seems able to resist it while, at the same time, not facing problems of reasons-responsiveness that affect blockage cases.

In this section, we shall try to rescue the "flicker" strategy from the main objection to it, namely the robustness objection. First, we shall address an important point in Fischer's original statement of the objection, namely that it is utterly mysterious how alternative pathways along which the agent does not act freely can make it the case that in the actual path the agent acts freely and is morally responsible for her action. Second, concerning Pereboom's example, we will resist Pereboom's and Zagzebski's contention that thinking of or attending to a moral reason is not the sort of factor whose presence or absence is (explanatorily) relevant to moral responsibility. On this basis, we shall argue in the next section that agents in plausible Frankfurt cases actually have robust alternatives even in Pereboom's sense. The conclusion will be that PAP is safe against Frankfurt-inspired attacks.

Let us start this programme by focusing on Fischer's "alchemy" charge against the forward-looking version of the "flicker" strategy. In the alternative pathways featured in Frankfurt cases, once the agent shows the triggering sign and the counterfactual factor takes over, the agent does not act freely and is not morally responsible for what she does. On this basis, Fischer's point is that obtaining freedom and moral responsibility in the actual sequence from their absence in the alternative sequence, as some "flicker" theorists apparently try to do, looks like "alchemy", and this clearly indicates the irrelevance of those alternative paths to the agent's freedom and responsibility. For moral responsibility, what counts is the actual causal history of the action. If the agent deliberated, decided and acted on her own, free from coercion or compulsion, she is morally responsible for what

she did, no matter whether she could have done otherwise. An alternative pathway along which the agent does not act freely and is not morally responsible for her act cannot plausibly be considered as a robust alternative, that is, an alternative that can ground her moral responsibility for what she actually does. This alternative pathway, available though it may be, is explanatorily irrelevant to moral responsibility.

Now, we may agree with Fischer that, when we think about an agent's having alternative possibilities, we most naturally imagine cases in which it is up to her to perform a different action, so that this alternative action, had she performed it, would also have been free, in the sense relevant to moral responsibility. But this does not mean that we do not also have pretty clear intuitions about different cases, in which the agent's only alternative pathway is to do unfreely and involuntarily what she actually does freely and voluntarily. In these cases, our judgement is that the fact that she did *not* take the alternative pathway, in which she acts unfreely and is not morally responsible for what she does, is explanatorily relevant to her moral responsibility for what she does in the actual path: she is morally responsible for what she did partly because she did not take that alternative pathway. A case of this sort follows. Let us call it "Activist1".

Imagine (no strenuous effort required) a country ruled by a dictatorial government, where some clandestine opposition groups operate. The government wants these groups to be discovered and dismantled. Police officers sometimes capture members of some of these groups and attempt to force them to reveal the identities of their comrades. In the past, the police used to employ torture, when necessary, for this purpose. Now, however, they do not, and the government is pleased about this, because it improves the regime's image in the eyes of foreign and domestic observers. The reason is not that they have become human rights champions, but that a research group has recently discovered a drug that, if taken by someone, makes it virtually impossible for her to lie. After taking it, the subject will say everything she knows about any question that she is asked. The truth drug (TD), as we may call it, has some harmful, though not fatal, side effects on the recipient's health. All this is by now common knowledge, especially among members of clandestine opposition groups. One day, Helen, a member of one of these groups, is arrested by the police during a political demonstration and taken to police headquarters. As is now usual, she is given a choice by police officers: she can voluntarily reveal the identities of her comrades; but, if she refuses, TD will be administered to her and she will give them the information anyway. Suppose that, fearing the harmful side effects of TD, and thinking that in any case the police will obtain the information they want from her, she chooses the first alternative and voluntarily and on her own reveals the information that, had she refused to do so, she would have revealed involuntarily and unintentionally.

A hard choice, to be sure, but sometimes life confronts us with hard choices. Now, my intuition is that Helen did something morally wrong, that she did so voluntarily and with all the freedom (not much, to be honest), including freedom to do otherwise, that was available to her in the circumstances, that the only alternative course of action she had was to do involuntarily and unfreely what she did willingly and freely,[11] that she is morally responsible for what she did, and that this moral responsibility *is, at least partly, explained* by the fact that she ought to, and could, have done otherwise, even if "otherwise" only means here to do *unfreely* what she actually did freely and on her own. I hope this intuition will be widely shared. If someone hesitates, it may help to imagine oneself in the shoes of one of Helen's comrades. Suppose that Helen said to one of them:

"Well, why should I wait for the TD to be administered? You know it has harmful effects and they would have had the information anyway. I had no alternative." If I were Helen's comrade, I would certainly not accept this plea as exempting her from moral responsibility. Helen ought to have taken the other alternative, and the fact that she did not take it partially explains our judgement that she is morally responsible, morally blameworthy, in fact, for what she actually did.

Now, contrary to Fischer's contention, this case shows that no mystery, no alchemy seems to be involved in the fact that pathways along which an agent does not act freely and is not morally responsible for what she does can be relevant to explaining why she acted freely and was morally responsible in the actual pathway. And, given that in Frankfurt examples it is also the case that, in the alternative pathways, after the device takes over, the agent does not act freely, and is not morally responsible for what she does, this significantly raises the prospects of forward-looking versions of "flicker" strategies to resist the "alchemy" charge. Helen had the sort of alternatives emphasized by some "flicker of freedom" theorists, namely to do something on one's own or not on one's own (Naylor), or to do something intentionally or unintentionally (Davidson). So this sort of alternative *can* be robust. And bear in mind that *they are structurally present in all classical Frankfurt cases,* including Pereboom's. Here, Joe can evade taxes freely and on his own or do it unfreely and not on his own. And it can now be argued that he is morally responsible for what he did partly because of the availability of that alternative pathway.

However, critics of the "flicker" strategy, such as Fischer, Pereboom and even Widerker, are unlikely to be impressed by Helen's example. They will certainly point out that, though Helen would not have acted freely after having the TD administered to her, her *choice* to have it administered would have been free and under her control, as was her actual choice. Widerker would contend that the example leaves Helen's control over her choice intact, and it is her freedom to choose otherwise that, at least partly, grounds her moral responsibility. Pereboom and Fischer would agree with Widerker that alternative free choices are robust alternatives and that Helen had them, but would argue that they are not required for moral responsibility, for it is possible to construct examples where alternatives of that kind are absent, the agent is morally responsible, and her only alternatives are not robust enough to ground her moral responsibility.

Now, examples such as Fischer's, where the alternatives are purely involuntary happenings which take place before the agent does anything freely, succumb to Widerker's dilemma. But an example such as Pereboom's seems immune to it. In examples of this kind, the agent has some control, even voluntary, over the alternatives, but these are not robust enough to explain her moral responsibility. Unlike Helen, it cannot be said that Joe, in Pereboom's example, could have *chosen* to evade taxes unfreely. To address this issue, let us modify Activist 1 so as to get it closer to Pereboom's case. We shall call this modified example "Activist2".

Suppose now that Helen is ignorant of the existence of the TD, though she has been told that the police do not use torture any more (but who knows!). Suppose further that Helen's psychology is similar to Joe's in Pereboom's example. In Helen's case, only a certain level of attention to moral reasons could lead her to take seriously the possibility of refusing to denounce her comrades (though she might still do so after taking such a possibility seriously). TD will be now administered to her as soon as she takes this possibility seriously as a consequence of her attending to moral reasons. To make the

case more realistic, suppose, as seems plausible, that her taking this possibility seriously gives rise to an emotional state that significantly increases her blood pressure, and that this causal correlation between thinking seriously of doing something that is felt as potentially dangerous and an increase in blood pressure is well known to psychologists and doctors. When she gets to police headquarters, she is examined by a doctor, so that the police can later defend themselves against possible charges of torture. During this examination, and unknown to her, the doctor leaves installed on her skin a sensor that detects an increase in her blood pressure, so that, as soon as this increase is detected, TD is immediately injected into her. However, when she is asked by the police officers to give them the identities of her comrades, Helen does not attain the required level of attention to moral reasons and does not seriously consider the possibility of refusing to comply with the request. TD is not administered to her and she voluntarily gives the police the information they want.

Helen had the following alternatives: either to attend to moral reasons against denouncing her comrades or not, and, as a result of them, to denounce her comrades, though not voluntarily,[12] or to denounce them voluntarily. Unlike the original example, now Helen cannot *choose* to denounce her comrades involuntarily. Now, against Widerker, I think that Helen was morally responsible for choosing to denounce her comrades even though no alternative choice was open to her. Alternative choices, then, do not seem necessary for moral responsibility for one's choices. However, against Pereboom and Fischer, I do not think that the alternative that Helen had, namely paying enough attention to moral reasons and taking the possibility of not denouncing her comrades seriously, as well as her denouncing them involuntarily, was explanatorily irrelevant to her moral responsibility for denouncing them on her own. Knowing all the objective facts, Helen's comrades would be justified in blaming her for what she actually did, namely for not attending sufficiently to moral reasons against denouncing them and for denouncing them on her own, on the basis that she ought to, and could, have done otherwise.

It is true that, in Activist2, the alternative of attending to moral reasons is not robust *in Pereboom's sense*. To see this, let us provisionally erase the counterfactually intervening factor, namely the TD injection. Suppose now that Helen attends to moral reasons in the required way and seriously considers the possibility of refusing to comply with the police's request, but that, none the less, she finally dismisses it and decides, on the basis of her self-interested reasons, to denounce her comrades. In this case, she obviously could not get rid of her responsibility for denouncing her comrades by claiming that she seriously attended to moral reasons against such an act. But this does not mean that this claim is completely irrelevant to her moral responsibility. In this case, it may increase that responsibility, for it shows that Helen's decision to denounce her comrades was not a sudden, unreflective reaction, but a reflective mental act. It is a mistake for Pereboom to think that an alternative pathway can only be relevant to an agent's moral responsibility and to assessments thereof if following it would *exempt* her from moral responsibility, for this assumes that moral responsibility does not come in degrees, which seems clearly false. It makes perfect sense to say of an agent that she bears more or less responsibility for a decision or an action of hers, depending on circumstances of various kinds. And these circumstances may include her attending to moral reasons. This factor may influence the moral responsibility assessment in several ways. Attending to moral reasons

may aggravate the assessment, as suggested, when the agent dismisses them, but again this need not always be the case. The agent may convincingly argue that these reasons were outweighed, in the particular case, by non-moral considerations, or even by contrary moral reasons. On the other hand, *not* taking moral reasons seriously can also modify the assessment in different ways in particular cases. It may temper the judgement about the agent's moral responsibility if, for instance, it was difficult to view the situation from a certain moral perspective. But it may also aggravate the judgement by presenting the agent as inconsiderate, selfish or uncaring. This seems to be the case in Activist2.

So whether an agent has properly attended to moral reasons or not is something we often take into account in order to ground our moral responsibility ascriptions. Against Zagzebski, who joins Pereboom in defending the irrelevance of Joe's alternative for his moral responsibility, it is simply not true that "in ordinary situations in which there is no machine in existence we do not blame [an agent] because she could have had such a [moral] thought… Having a moral thought is not the kind of power relevant to her blameworthiness" (Zagzebski 2000a:242). On the contrary, in ordinary situations it is quite common to blame an agent and to worsen our assessment of an action of hers for not having properly thought of or attended to moral reasons. Think, for instance, of the following remark, which clearly worsens the moral assessment of the agent: "How could you not care about the harm you would cause by doing that?" We rightly expect morally responsible agents to properly attend to moral reasons in situations that plainly have moral aspects and implications.

Let us go back to Pereboom's original example. The impression that Joe's alternative is explanatorily irrelevant to his moral responsibility may derive from the fact that the moral import and worth of Joe's act is not straightforward. Paying taxes is not self-evidently a morally good action. There are, for instance, tax-objectors, who refuse to pay taxes on the ground that they are not rightly spent by the government. Cases of corruption and embezzlement, as well as widespread attempts to evade taxes by hiding earnings from the Treasury's eyes, may also lead some morally honest people to feel that the tax collection system is not fair. In this context, consider one way in which Joe could try to justify his claim for an illegal tax deduction. He could say: "Well, I tried hard, but could not think of any decisive moral reason not to do it", and ground this assertion in arguments such as those above. And, in some circumstances, some people might accept this claim as exempting Joe from moral responsibility, or at least as significantly diminishing it. In some cases, forceful moral reasons may not be easy to find.

Now, if we think of a case where the agent's act merits a straightforward and widely shared moral assessment, the relevance for moral responsibility of such available alternatives as attending to moral reasons can be seen clearly. Activist2 may be such a case. But also imagine modifying Pereboom's example by replacing Joe by Jones, the neuroscientist by Black, and Joe's choosing to evade taxes by Jones's choosing to kill Smith. As in the original example, the only thing that might lead the agent to a different choice was her attaining a certain level of attentiveness to moral reasons against killing Smith. Now, in this modified Pereboom case, most people (leaving aside moral nihilists) will simply not accept Jones's claim that he tried but could not find any decisive reason not to kill Smith as exempting him from, or mitigating, his moral responsibility for his decision to kill Smith. The likely response will be that there are just too many decisive moral reasons against that decision, and that he ought to have paid proper attention to

them. This clearly implies that the fact that Jones did not properly attend to moral reasons is, in this case, both morally relevant and explanatorily relevant for his moral responsibility. He, like Helen, did not exhaust the possibilities that were open to him; had he exhausted them, he would still have decided to kill Smith and done so (given Black's intervention), but he would not have been morally responsible for it. And part of the reason is that then, through no fault of his own, he could not have done otherwise.

Robust (exempting) alternatives in Frankfurt cases

Some responses to the preceding line of argument are not hard to anticipate. First, in the modified Pereboom example, Jones is morally responsible for choosing to kill Smith and for doing so, but he could not have avoided *that*. Something similar could be said about Activist2. Second, if we erase Black from the picture, even if Jones had done the only thing he could freely do when Black was present, namely attending properly to moral reasons not to kill Smith, this would not have precluded his moral responsibility if he, none the less, had finally decided to kill Smith and done so. But this shows that the alternatives that were available to Jones (or Helen, in Activist2) do not satisfy Pereboom's strong criterion of robustness, namely that, by taking them, the agent would, and understood that she would, be *precluded* from her moral responsibility. Let us call robust alternatives in Pereboom's sense "exempting alternatives". So, the objector could admit that we have shown that alternatives indicated by "flicker" theorists in Frankfurt cases can have a certain kind of robustness or explanatory relevance for moral responsibility ascriptions and assessments, but still insist that we have not shown that they are robust in Pereboom's "exempting" sense.

Let us deal with the second objection first. The objection amounts to the following: not merely attending to a moral reason, but only *deciding and acting on that reason* could be an exempting alternative. Only alternative decisions and actions can actually be exempting alternatives. But these alternatives were not present in Activist2 or in Pereboom's (original or modified) example. If the agents are still responsible in those cases, it is not because they had exempting alternatives, for they did not have any. So exempting alternatives are not required for moral responsibility. As for weaker alternatives, they may modify moral responsibility assessments, but do not ultimately explain why the agent is morally responsible.

Our response starts by taking the objection to its ultimate consequences. The objection is, in fact, too concessive in allowing alternative *decisions* to be exempting alternatives. If, in the spirit of Pereboom's proposal, an exempting alternative has to be such that, if the agent took it, she would (and she understood she would) thereby be precluded from her moral responsibility for what she actually did, then decisions do not meet this condition. For the agent might decide to perform an alternative action and, when the time comes, change her mind and not perform it. In this case, she obviously could not avoid her responsibility for what she did by claiming that she had decided to do something different. In fact, the condition is not met by Pereboom's willings either. Remember his characterization of a robust (exempting) alternative: the agent "could have willed something different from what she actually willed such that she understood that by willing it she would thereby be precluded from moral responsibility for the action" (Pereboom 2001:26). Remember Activist1. Imagine that Helen actually wills not to

reveal her comrades' identities on her own and therefore she wills the Truth Drug to be administered to her. But suppose that, when she sees the syringe, she gets scared and denounces her comrades on her own. She obviously could not avoid (even though she could lessen) her responsibility for denouncing her comrades on her own by claiming that she had willed something different. The only exempting alternative she had was *to act on that decision* by having the drug administered to her. We have said that *only acting on a reason, and not merely attending to it, can be an exempting alternative*. But we can now add, in the same vein, that *only acting on a decision, and not merely making it, can be an exempting alternative*. This suggests that, for moral responsibility, decisions may not play so crucial a role as many thinkers (including Widerker and Pereboom) have supposed they do.

But now it seems that, even in the simplest Frankfurt cases, agents do not have exempting alternatives. They may have the sort of less robust alternatives indicated by "flicker" theorists. These alternatives, we have argued, are more relevant to ascriptions of moral responsibility than some authors, such as Pereboom, have supposed. However, they are not, as such, exempting alternatives. Even if an agent, in a Frankfurt case, can show an inclination towards an alternative way of acting, or attend to a reason for it, or even decide to follow it, these are not, as such, exempting alternatives. Only acting on such an inclination, or reason, or decision, would be. But, because of the counterfactual intervener, this is precisely what agents in Frankfurt cases seem unable to do. However, if we look carefully at the examples, this inability might emerge as an appearance. Let us argue for this.

As a preliminary, though important, remark, we should note that "S is morally responsible for…", as well as "S did intentionally…", and related locutions, introduce intensional contexts. Even if "A" and "B" are different descriptions of the same behaviour, an agent can be morally responsible for A and not for B, and she can intentionally do A and not B. Moreover, some kinds of actions can only be performed intentionally.[13] This is the case with lying, murdering someone, denouncing someone, and many others. Keeping this in mind, let us go into the examples.

Think first of Activist1. As this example shows, the fact that, in the alternative sequence, an agent does not act freely and intentionally does not mean that this cannot be a robust, exempting alternative. Helen has such an alternative: she could have acted on a decision she could have made, namely the decision not to reveal the identities of her comrades voluntarily. Had she acted on that decision, she would not have been morally responsible for what she actually did. But we can go a bit further. It would be wrong to say that, in the alternative sequence, she still denounces her comrades, though not intentionally or voluntarily. She gives the police their identities, but does not denounce them, for denouncing someone is not something one can do unintentionally. In the light of this, Frankfurt seems too rash when, in his original paper, he writes about Jones that "whether he finally acts on his own or as a result of Black's intervention, he performs the same action" (Frankfurt 1969:8). The assumption that, in the alternative sequence of Frankfurt cases, an agent performs the same (kind of) action as in the actual sequence is widespread in the literature,[14] but it is wrong.

Now think of Activist2. In this example, Helen could have attended to moral reasons, but did not. Though morally and explanatorily relevant, this is not, as such, an exempting alternative. Only acting on those reasons would be. Initially, it does not seem that

Helen could have acted on those reasons. But, again, this is only an appearance. The moral reasons Helen could have properly attended to were reasons *not to denounce her comrades*. Had she acted on those reasons, she would have avoided her moral responsibility for denouncing them, and she clearly understood that this would be the case. But *she could have acted on those reasons,* for, had she attended to them, the TD would immediately have been administered to her. As a result, she would have revealed her comrades' identities, but, as this would have been unintentional, she would not have *denounced* them. She could, then, have acted on the reasons she ought to, and could, have attended to. So Helen had exempting alternatives. Something similar can be said about Pereboom's modified example, where Joe is replaced by Jones and the evasion of taxes by the killing of Smith. Here, the moral reasons Jones could have properly attended to were reasons not to *murder* Smith. Acting on those reasons would have precluded his moral responsibility for murdering Smith, and Jones clearly understood that this would have been the case. And he could have acted on those reasons. In the alternative sequence, Jones's behaviour, in the sense of his physical movements, might have been the same as in the actual sequence: he might physically cause Smith's death as he does in the actual sequence. However, and this is the crucial point, in the alternative sequence, unlike the actual one, Jones does not *murder* Smith. It is not that, in the alternative sequence, Jones murders Smith, only not on his own. He does not murder Smith at all. He may cause Smith's death, but this is a different kind of action than murdering Smith. In examples with less pressing moral profiles, such as Pereboom's original example of the tax evader, we may not have different terms to apply to what the agent does in the actual and in the alternative sequences, but we clearly have the concepts: evading taxes intentionally and on one's own is different from doing so as an effect of a device's firing in one's brain.

The standard objection against the "flicker" strategy, as we have seen, is that agents lack control over the alternatives. This may be true in cases in which the alternatives are mere physiological happenings lacking intentional content, such as blushes or twitches. But, as we have argued, these cases are not convincing. On the one hand, they fall prey to the dilemma defence. On the other hand, they raise the suspicion that the agent does not reach her decision through a normal process of deliberation, which in turn erodes the judgement about her moral responsibility. Blockage cases, we have argued, are also affected by the latter point, and may also be affected by the former. However, in Frankfurt cases that are immune to these problems, such as Pereboom's, it is far from clear that the alternatives are beyond the agent's control. We ordinarily assume that agents ought to, and could, attend to certain reasons. If it is not reasonable to expect that the agent could think of or attend to these reasons, this certainly affects our assessment of her moral responsibility in particular cases. In extreme cases of "morally blind" agents, ascriptions of moral responsibility may lose their ground.

Moreover, the connection between alternatives of this sort and what takes place after the agent takes them is far from accidental. Even if, as a result of the agent's taking them, she acts unintentionally, an internal connection is preserved. In Activist2, or in Pereboom's original or modified examples, in the alternative sequence the agent in fact acts on the reasons she ought to, and could, have attended to, even if her so acting is unintentional. And, as Activist1 shows, the fact that, in an alternative sequence, the agent acts unintentionally can be clearly relevant to the moral responsibility she bears in the

actual sequence. There is a deep explanation of the non-accidental character, in the alternative sequence of the examples mentioned, of the link between an agent's attending to a moral reason and her acting on that reason, namely that, on any plausible view, a reason, unlike, say, a blushing or a twitching, conceptually includes the notion of the action for which it is a reason. And the same holds for decisions. In Activist2, for instance, the moral reason Helen attends to in the alternative sequence is a reason for *not denouncing her comrades*. So, even though, in this case, she cannot freely decide and act on that reason, the reason still keeps a non-accidental relation to her behaviour in the alternative sequence, for, as we have seen, in that sequence she does not denounce her comrades. So she was morally responsible for denouncing her comrades, but it was within her power not to have denounced them, by virtue of it being in her power to attend to moral reasons against doing so. And the case can be generalized to those Frankfurt cases that do not fall prey to the dilemma defence or to suspicions about the rationality of the process leading to the agent's decision and action. PAP, then, is thereby vindicated.

Let us now address the first objection we mentioned. This is, in fact, the objection that Kane raised against Naylor's response to Frankfurt cases. Naylor holds that, in the Jones/Smith/Black case, Jones is morally responsible for killing Smith on his own, something to which he has alternatives, but not for killing Smith, which he has no alternative to. As we saw, Kane rightly criticized this move by pointing out that it "too artificially separates responsibility for doing-A-on-one's-own from responsibility for doing A. In general, if we are responsible for doing something on our own, we are responsible for doing it" (Kane 1996:41). As we suggested, a similar criticism could be raised against Davidson's proposal, by substituting "intentionally" for "on one's own" in Kane's quotation. Now, our proposal has a close relation to Naylor's (and Davidson's) moves. Remember our modified Pereboom example. If, in the alternative sequence, unlike the actual one, Jones does not murder Smith, this is partly because he does not bring about Smith's death intentionally and on his own. It seems, then, that a parallel criticism would affect our proposal as well if, following Naylor's steps, we also hold that Jones is responsible for murdering Smith, but not for killing him or for bringing about his death. And, *mutatis mutandis,* something similar holds for Helen in Activist 1 and Activist2.

This conclusion, however, is not forced upon us by a defence of PAP. There are logical and hierarchical relations between certain action-descriptions and the corresponding action types. This is the case with "A murders B", "A kills B", and "A brings about (or causes) B's death". Each member of this chain is conceptually sufficient, but not necessary, for the next. And each member is conceptually necessary, but not sufficient, for the preceding one. So, if A murders B, then it has to be true that A kills B and that A brings about B's death, but the truth of the latter does not imply the truth of the former. I can kill someone accidentally, and thereby bring about his death, but this does not mean that I have murdered him. Though some logical relations are far from evident and need a careful reflection to be seen, the logical and hierarchical relations we have described seem commonplace and are almost immediately accepted when proposed. On this basis, it seems plausible to say that there is also a corresponding logical and hierarchical chain among responsibility attributions involving those action descriptions. So it seems true to say that, if A is responsible for murdering B, he is also responsible for killing B and bringing about B's death, since these are, and the agent knows they are,

plainly conceptually necessary for the former. Since I cannot (and know I cannot) murder someone without killing him and causing his death, I am responsible for the latter by being responsible for the former. But the converse is not true. Since I can kill someone or cause his death without murdering him, I can be responsible for the former without being responsible for the latter. Responsibility, we might say, transmits itself through one of these chains in a top-down, not bottom-up, direction.

If this is correct, we can plausibly meet Kane's objection. "Killing Smith" and "causing Smith's death" are conceptually implied by (conceptually necessary for) "murdering Smith". Then, if we accept that Jones is morally responsible for murdering Smith, to which he had alternatives, PAP does not require us to deny that Jones is also morally responsible for actions that were (and Jones knew they were) conceptually required for his murdering Smith. In more technical terms: if Jones is responsible for his action to be truly described as "murdering Smith", and this cannot (and Jones knows it cannot) be truly so described without also being truly describable as "killing Smith" and "causing Smith's death", then he can be morally responsible for the latter in being responsible for the former. In the actual sequence, the descriptions of Jones's action as "killing Smith" and "causing his death" are true *because* they are conceptually necessary for the description "murdering Smith" to be true, but this does not apply to the alternative sequence. In less technical terms: in the actual sequence, but not in the alternative sequence, Jones kills Smith and causes his death as necessary constituents of his murdering Smith. In the same vein, we can also say, without violating PAP, what seems natural and true to say, namely that Jones is morally responsible for Smith's death, since, in the actual sequence, but not in the alternative sequence, Smith's death is (and Jones knows it is) brought about as a conceptually necessary constituent of Jones's murdering him. Again, *mutatis mutandis,* something similar applies to Helen in Activist1 and Activist2, and to Joe, the tax evader, in Pereboom's original example.

Let us finally address a worry that someone may feel about our preceding defence of PAP. Remember that Frankfurt cases are intended to prove that we do not actually assume that agents must have alternative possibilities in order to hold them morally responsible for their actions. To achieve this result, they have to confront us with cases in which we intuitively ascribe moral responsibility to an agent in spite of the fact that she did not have alternatives open to her, or at least alternatives that are robust enough to account for her moral responsibility. We have tried to show that in Frankfurt cases agents *do* have robust alternative possibilities open to them. But it might still be objected that we have not shown that these alternatives are taken into account by us when we intuitively form the judgement that the agent is morally responsible. The objection could point to the fact that a good deal of philosophical reflection has been needed in order to argue for the presence of robust alternatives in Frankfurt cases, and that it is not plausible to hold that the intuitive, pre-theoretical judgement that the agent, e.g. Jones, is morally responsible rests upon such philosophical technicalities. So even if Jones does in fact have robust alternatives, this might still not be what accounts for our judgement that he is morally responsible for his action. If so, we have not yet shown that our everyday concept of moral responsibility honours PAP.

In response, we have to agree that seeing that Jones had the alternatives of murdering and not murdering Smith (or that Helen had the alternatives of denouncing and not denouncing her comrades) does require complex philosophical reflection and that it is not

plausible to expect this reflection to inform our intuitive responsibility ascription. However, this reflection only articulates and gives a theoretical grounding to what we intuitively, pre-theoretically may feel in confronting Frankfurt cases, namely that, though in some rather neutral sense the agent is doing the same thing in both the actual and the alternative sequences, in another, and morally relevant, sense, she is not, and that this difference matters to our moral judgement. Philosophical reflection does not create this difference, but only gives it a theoretical and more articulate form.

I think we are now allowed to conclude that PAP is safe against Frankfurt-inspired criticisms. This principle, however, has also been challenged from other perspectives, which rely on two other sorts of cases. We may call them "self-trapping" and "Luther cases", respectively. Let us turn to them.

"Self-trapping" cases

Sometimes agents put themselves freely into a position in which they cannot do otherwise with respect to an action or omission which, none the less, they can rightly be held responsible for. Think of the following example, devised by James Lamb:

> Peggy Sue might have got herself arrested so that she would not have to appear at some dull function she had promised to attend. It would seem that she is morally responsible for having broken her promise—we would certainly be justified in faulting her—and yet it would also seem that, being under lock and key, she could not have kept her promise to attend the party.
>
> (Lamb 1993:518)

This sort of example does not require us to abandon PAP. It is rather obvious that Peggy Sue could have attended the party and not have broken her promise. Not, to be sure, after being arrested, but, since she put herself freely in a situation which she knew would make it impossible for her to attend and she could have avoided putting herself in that situation, it seems simply right to say that she could have attended the party. In its usual formulation, PAP does not require an agent to be able to act otherwise at the very time at which she acts. Ekstrom disagrees. She holds that the most natural way to understand PAP is as follows: "PAP1. A person is morally responsible for doing X at t only if s/he could have done otherwise than X at t" (Ekstrom 2000:200). And she goes on to hold that, though Frankfurt cases do not falsify PAP, self-trapping cases do. But even if it is natural to read PAP as PAP1, this need not be the only correct way of reading it. If PAP1 is falsified by self-trapping cases, PAP itself need not be, for PAP itself does not specify the time at which the agent has to be able to do otherwise, and it allows its implicit temporal index to be specified in several ways. In one of these ways that PAP leaves open, it is simply true that Peggy Sue could have attended the party.

The preceding case does not constitute a powerful obstacle to PAP defenders. But other cases may be harder to deal with. In Peggy Sue's case, it is quite obvious why she intentionally got herself arrested. We can easily construe a (Davidsonian) reasons explanation of her way of acting: she did not want to attend the party and she believed (rightly, as it turns out) that acting in such a way that she got arrested would allow her to

satisfy her desire (while, at the same time, providing her with an excuse for having broken her promise); and, because of that reason, she did act in such a way. In teleological terms: she got herself arrested in order not to (be able to) attend the party (and to have, at the same time, a credible excuse for breaking her promise to attend). The alternative pathway, in which she does not act so as to be arrested, had she taken it, would also have had a rational explanation, in terms of her willing to keep her promise, in spite of her reluctance to attend the party. But think of this other case, which Lamb takes from A.S.Kaufman:

Suppose that a lifeguard who has lied about her qualifications is unable to swim. Assume now that a child drowns whose life it was the lifeguard's duty to save. We would certainly hold the lifeguard responsible and yet, being unable to swim, she could not have saved the child's life.

(Lamb 1993:525)

Though Lamb does not classify this example under the same category as Peggy Sue's, we think (though nothing important hangs upon this) that it also deserves to be called a case of self-trapping. There is also something the lifeguard freely and voluntarily did which put her in a situation in which she did something for which she is morally responsible even though she could not have done otherwise. Or so it seems. The two cases are, however, interestingly different, and not only because of the graver moral significance of the latter. Peggy Sue clearly got herself arrested in order not to (be able to) attend the party, but the lifeguard did not lie about her qualifications (unless she was a moral monster) in order not to (be able to) save the child's life. Not saving the child's life (unlike not attending the party in Peggy Sue's case) was not something she wanted or intended to bring about with her lying. And I certainly share Lamb's judgement that the lifeguard is morally responsible for not saving the child's life, even if, being unable to swim, she could not have done otherwise.

Lamb thinks this case is in fact a counterexample to PAP, even if we allow for several specifications concerning the time of the alternative action. His reason is that there is (or at least there need be) no time at which the lifeguard could have saved the child, whereas, in Peggy Sue's case, there is a time when she could have attended the party. He writes:

We may suppose, for example, that the lifeguard is unable to swim, not because she never happened to learn, but because of some irresistible genetic trait, such as an inborn fear of the water, which made it impossible for her to learn. It can thus be supposed that the lifeguard at no point in her life could have done anything to save the child's life.

(Lamb 1993:525–6)

He then proposes a formulation of a related principle which is safe against this and previously considered counterexamples, such as Peggy Sue's. He calls this the "weak principle of alternate possibilities". According to this principle, "a person is morally responsible for doing something only if at some time there is *something* he could have avoided doing" (Lamb 1993: 527). It is pretty clear what Lamb is intending to capture with this principle. In the case at hand, there is something the lifeguard could have avoided doing, namely lying about her qualifications. And the relevance of this act,

which she ought to and could have avoided, for her moral responsibility for failing to save the child is undoubted. But Lamb's principle does not contain any reference to the relevance of the alternative action for the agent's moral responsibility, so that an alternative pathway with no connection at all with that for which the agent is morally responsible might satisfy the condition it states. A lesson to be drawn from the robustness objection is that any plausible alternative possibility that is proposed as a necessary condition of an agent's moral responsibility for a certain action has to be sufficiently robust. The alternative has to play a role in grounding and explaining that moral responsibility. The spirit of Pereboom's proposal for spelling out this role is that the alternative has to be such that, had the agent taken it, she would thereby have avoided her moral responsibility, and also such that she understood that this would be the case. In fact, we went beyond Pereboom's contention that alternative willings may meet these requirements and argued that only alternative actions may in fact meet them. Remember that we held that, in Frankfurt cases, agents had robust alternatives in this sense. In the Jones/Smith/Black case, for instance, Jones had such an alternative, namely not to murder Smith.

Now, the literal formulation of Lamb's principle does not respect these conditions, but it might be enlarged so as to respect them. So a plausible principle to cope with the lifeguard's case might be the following: a person is morally responsible for doing something only if at some time there is something she could have avoided doing such that, had she avoided doing it, she thereby would, and she understood she would, have precluded her moral responsibility for what she did. Does the lifeguard satisfy this condition? She is responsible for not saving the child's life, that is, for an omission or failure to do something. Unlike Jones, she did not want or intend that for which she is morally responsible. Her failure to save the child is an unwanted and unintended consequence of something she freely and intentionally did, namely lying about her qualifications, which she could have avoided doing. But there is a rationale for holding her responsible for this consequence, namely that this consequence, even if unwanted and unintended, can be expected to be foreseen by an agent with normal cognitive powers as likely to arise from an act like the one performed by the lifeguard. Now she ought to, and could, have foreseen this possible consequence and she ought to, and could, have avoided it by not lying about her qualifications.

Now, if the spirit of PAP, even robustly interpreted, is preserved by the amended Lamb's principle, it may seem that the letter is not. For it is still the case that the lifeguard is morally responsible for not saving the child's life even though she could not have saved it.

But in fact there is nothing in these considerations that falsifies PAP itself, robustly interpreted. The lifeguard satisfies the condition imposed on being responsible by Lamb's amended principle, but I think she also satisfies the stronger condition imposed by PAP itself. Let us argue for this contention.

Both Lamb's principle and its amended version are stated in terms of responsibility for actions ("…a person is morally responsible for *doing* something…"). But, in fact, in the lifeguard example we are specifically concerned with responsibility for *not doing* something. The lifeguard is morally responsible for *not* saving the child. And the issue of responsibility for not doing something or, as is usually expressed, responsibility for omissions, has some special complications. Consider that if, at a certain time, we can

truly be said to be doing just one or at most a few things there are instead millions of things that we are not doing. There are two related but distinct ways in which a locution of the form "S does (or did) not A" can be understood. In one of them, it states what it literally means: that S does not perform an action of a certain kind. But this literal meaning does not exhaust its content, especially in the context of responsibility attributions. Normally, when we say that S is responsible for not A-ing, we do not simply imply that S did not A, but also that doing A was expected or required of him. According to the Oxford Dictionary of Current English, "non-performance of what is normal, expected or required" is one of the meanings of the word "failure", and this is how we shall understand the word in what follows. There are higher and lower degrees in which doing something is expected or required of a person at a particular time and place, as well as corresponding degrees in the responsibility she bears for her failure to do it. For example, a professional firefighter is more strongly required to save a person from a fire than an ordinary person, and her degree of responsibility for her not doing so is correspondingly higher. Let us call a firefighter's not saving someone from a fire when she could do it, as well as similar cases, such as a lifeguard's not saving a drowning person, *specific failures,* while in the case of an ordinary person we may call such omissions simply *failures*. Finally, someone may not do something without this being either a failure or a specific failure. There are thousands of examples of this. A person unable to swim does not commit a failure in not saving a drowning person. But more ordinary cases also fall under this category. My not reading a novel or not drinking water now are easy examples.

Now, "S specifically fails to A", "S fails to A" and "S does not A" relate to each other in a logical and hierarchical way quite similar to the way in which we argued above that "A murders B", "A kills B" and "A causes B's death" do. So S's not A-ing is a conceptually necessary, but not a sufficient, condition for S's failure to A, while this is in turn a conceptually necessary, but not sufficient, condition for S's specific failure to A. Each member of this chain implies the previous one, but is not implied by it. So not every case of not A-ing is a failure to A, nor is every failure to A a specific failure to A, whereas every specific failure to A is a failure to A and every failure to A is a case of not A-ing. As another example, I do not teach chemistry, but this is simply something I do not do; it is clearly not a failure of mine, since I am not expected or required to do it, nor is it a specific failure of mine either. But a professional chemistry teacher might be committing a specific failure, and therefore a failure, in not teaching her chemistry classes.

On this ground, it seems simply true to say that the lifeguard's not saving the child is a specific failure (ordinary people at the beach might be committing a failure, but not a specific one, in not saving the child, while people unable to swim are not committing either). Now, as happens in the murdering/killing/causing-death chain, responsibility also transmits itself in the specific-failure/failure/not-doing chain in a top-down (or left-right) but not in a bottom-up (or right-left) direction, according to the logical relations between its members. Now, in the same way in which, in Frankfurt's case, Jones's responsibility for killing Smith, and for causing his death, is derivative, through these logical and hierarchical relationships, with respect to his responsibility for murdering him, the lifeguard's responsibility for not saving the child, and for her failure to do it, is also derivative with respect to her responsibility for her specific failure to do it. It would be an

understatement (though not a falsity) to describe Jones's case as one in which he is responsible for killing Smith or for causing his death, for this is consistent with his not being responsible for murdering Smith. And it would also be an understatement (but again not a falsity) to describe the lifeguard's case as one in which she is responsible for not saving the child, for this is consistent with her not being responsible for her failure, or for her specific failure, to do so. She is in fact morally responsible for not saving the child, but her responsibility for this omission derives from, and is owing to, her responsibility for her (specific) failure to do it. Bear in mind that other people present on the beach, but unable to swim, are not even responsible for not saving the child.

If this is on the right lines, then PAP is not falsified by the lifeguard example. The lifeguard is morally responsible for her specific failure, and for her failure, to save the child, but she could have done otherwise: she could have avoided committing those failures. By lying about her qualifications, which she could have avoided, she put herself freely in a position in which it was specifically expected and required of her to save people's lives, and so in a position in which her not saving a drowning person would become a specific failure, and not simply a failure or a simple case of not doing something. So she is morally responsible for her (specific) failure to save the child, but she could have done otherwise. As we have already suggested, we need not deny the intuition that she is morally responsible for not saving the child. She is, for her (specific) failure to save the child conceptually implies, and she knows it does, her not saving the child. Her responsibility for the latter, however, derives from, and is explained by, her responsibility for the former, with respect to which she could have done otherwise.

It is not by mere luck or accident that she was responsible for her failure. She knew that, and how, she could have avoided being responsible for it, and she did not take that alternative pathway, though she was able to take it. There is a probability relation that she, being a person with normal cognitive capacities, should, and could, have perceived between her lying about her qualifications and her possible (specific) failure to save a drowning person's life. So she had an alternative pathway open to her such that, had she taken it, she would have precluded herself from moral responsibility for her (specific) failure to save the child's life, as well as for not having saved it, and she knew that this would have been the case. On this alternative pathway, her responsibility for not saving the child's life would have been the same as that of other people present on the beach but unable to swim, namely none. The alternative possibilities the lifeguard had were, then, undeniably robust. So it seems that we can conclude that, even on a robust reading of "could have done otherwise", PAP is not falsified by the lifeguard example.

This result may have additional significance if we consider that the lifeguard case might also seem to threaten the principle that "ought" implies "can" (OIC). If we accept that the lifeguard ought (had a moral obligation) to save the child's life, but deny that she could do that, we are denying that principle. But a reasoning parallel to the one we followed in order to show that PAP is not falsified by the lifeguard example, with "moral obligation" substituted for "moral responsibility", would also yield the result that the OIC principle is not falsified by that example either. Note that people present on the beach but unable to swim had no moral obligation to save the child, since they could not

do so, as the principle dictates. The lifeguard could have been one of them. And this shows that her moral obligation to save the child derives from her moral obligation not to (specifically) fail to save the child, which she could have discharged by not lying about her qualifications.

"Luther" cases

Up to now we have been dealing, almost exclusively, with cases of blameworthiness, cases in which an agent is morally responsible for doing something morally wrong. In fact, when we use the locution "…is morally responsible for…" we normally do so to blame someone. This fact in itself deserves an investigation that cannot be pursued here. It seems to reveal a rather dark human propensity to devalue others, instead of searching for their valuable traits and performances. But this fact may also have less depressing explanations, which derive from the structure of our moral responsibility concept. Some philosophers, most prominently Susan Wolf (cf. Wolf 1990), contend that there is an asymmetry, concerning freedom and moral responsibility, between blameworthy and praiseworthy actions. In Wolf's view, which she calls the Reason View, freedom, as required for moral responsibility, essentially involves an ability to act on one's values and to form them out of an appreciation of the True and the Good. On this basis, she contends that an agent is blameworthy only if she has the required freedom and, therefore, only if she could have acted in a morally right way, but an agent can be praiseworthy even if she could not have acted in a morally wrong way, for in acting in a morally right way she already shows she has the freedom required for moral responsibility. In a more schematic way, alternative possibilities are required for blameworthy actions, but not for praiseworthy ones. Other philosophers, most prominently Daniel Dennett (cf. Dennett 1984), contend, on the basis that praiseworthy actions do not require alternative possibilities, that blameworthy actions, and in fact any actions for which an agent is morally responsible, do not require them, either, so that PAP is false.

In any case, however, both Wolf and Dennett agree that agents can be not only blameworthy (for their morally wrong actions), but also praiseworthy (for their morally right ones). "Luther" cases are of the latter sort, and they purportedly show that PAP is either partially (Wolf) or totally (Dennett) false.

We have chosen this label owing to Dennett's reference to Luther in the course of his criticism of PAP, but many other examples can plausibly fall into this category. As happened with "self-trapping" cases, here there are also softer and stronger examples. The strong ones, which include the Luther case, are especially challenging, and they reach far beyond the issue of alternative possibilities. Luther cases clearly merit more attention than has been paid to them, even in recent works on free will and moral responsibility such as Ekstrom's or Pereboom's. Let us start with a soft case, which we owe to Wolf:

Let us consider the woman who buys a gift for her friend claiming that she could not resist. Walking past a shop window, she sees a book that she knows her friend has been searching for for ages. It is only ten dollars, and so, imagining the delight on her friend's face when she delivers the book, she walks into the shop and buys it. Now there is

nothing in this story that tells us whether the woman could have refrained from buying the book, and when she says to her friend, "I couldn't resist," there is no reason to think that she means this phrase literally. After all, the woman didn't *try* to resist—how should she know whether she could have? Still, according to the Reason View, even if the woman's remark, taken literally, were true, this would not prevent her from being a responsible agent. Assuming that she did the right thing for the right reasons…the woman was responsible, and so deserves praise for her act of generosity whether she literally could have resisted performing it or not.

(Wolf 1990:83–4)

Wolf does not commit herself to the view that the woman in the example was saying something literally true in claiming that she could not resist buying the gift for her friend. In fact, it is difficult to accept that her claim was literally true. As Ekstrom points out, "such phrases are ordinarily taken as exaggerations in the name of humility" (Ekstrom 2000:165). What Wolf contends is that it would not matter for the woman's moral responsibility, and so for her praiseworthiness, if the claim is literally true or not. She would be praiseworthy in either case. Alternative possibilities are not required for the praiseworthiness variety of moral responsibility.

The example as such, however, does not establish this thesis. It would only establish it if it were literally true that she could not do otherwise than buy the book and, none the less, she would still deserve praise. But if we imagine the woman actually and literally satisfying that condition, our intuitions about her praiseworthiness will be much less firm, to say the least. If I were the woman's friend, I would be grateful if I knew she did not have any hesitations about buying the book, because this would be a sign of her love and friendship, but I would not be assuming thereby that it was literally *impossible* for her not to buy the book. If I came to suspect this, my attitude might be affected, precisely because it would lead me to doubt whether she really bought the present because of the reasons, namely her love and friendship towards *me,* which aroused my gratitude in the first place. I might start to see her love and friendship as forms of obsession, not as sound feelings. It is not clear, then, that the example shows that being unable to do otherwise does not matter for praiseworthiness.

Wolf's example, even if rather soft and clearly inconclusive, gives us a feeling of how "Luther" cases should be devised in order plausibly to question the necessity of alternative possibilities for moral responsibility. These cases have to feature a rational and morally sensitive agent who, at a certain time, sees a way of acting so clearly and strongly supported by her reasons and values, moral and otherwise, that any alternative pathway is simply discarded by her from the very beginning. These cases may invite us to see the agent, not merely as morally praiseworthy, but even as more so than if she had hesitated and taken some alternative way of acting seriously. Dennett's Luther example is of this sort:

"Here I stand," Luther said. "I can do no other." Luther claimed that he could do no other, that his conscience made it *impossible* for him to recant. He might, of course, have been wrong, or have been deliberately overstating the truth. But even if he was—perhaps especially if he was—his declaration is testimony to the fact that we simply do not

exempt someone from blame or praise for an act because we think he could do no other. Whatever Luther was doing, he was not trying to duck responsibility.

There are cases where the claim "I can do no other" is an avowal of frailty… I can do no other, I claim, because my rational control faculty is impaired. But in other cases, like Luther's, when I say I cannot do otherwise I mean I cannot because I see so clearly what the situation is and because my rational control faculty is *not* impaired. It is too obvious what to do; reason dictates it; I would have to be mad to do otherwise, and since I happen not to be mad, I cannot do otherwise.

(Dennett 1984:133)

On Wolf's behalf, let us say that Dennett is too rash in his appreciation of what the Luther example shows. It may show that we do not exempt someone from *praise* because we think he could do no other, but not that we do not exempt him from *blame*. There may be the asymmetry Wolf holds there is between praiseworthy and blameworthy actions with respect to the alternative possibilities condition and, if so, Luther's case cannot be held to show that this condition is not required for moral responsibility in general. This point is reinforced by the nature of the next example used by Dennett to defend his contention: "I hope it is true—and think it very likely is true—that it would be impossible to induce me to torture an innocent person by offering me a thousand dollars" (Dennett 1984:133). Again, this is supposed to be a case of someone's being praiseworthy.

That Luther cases may not affect the validity of PAP for what concerns morally wrong and blameworthy actions can be made more plausible if we think of cases in which it is literally true that an agent cannot do what is morally required. We have tried to show that neither Frankfurt cases nor "self-trapping" ones are of this sort, but, in discussing them, we have met some situations that are. Think for instance, in connection with the lifeguard example, of someone other than the lifeguard herself who, being unable to swim, does not save a drowning child's life. Though this (omission) is morally wrong, this person is not blameworthy for it, and clearly the reason is that she could not have saved the child.

Now, if Luther cases do not falsify PAP in connection with morally wrong and blameworthy actions, this is already an important point, for we have argued at different places that the falsity of PAP may imply the falsity of another principle, which in fact might partly explain the significance that PAP itself has for us, namely the principle that "ought" implies "can" (OIC). We can see now that only the falsity of PAP *in connection with morally wrong and blameworthy actions* might involve the falsity of OIC, for, even if PAP is false in connection with morally right and praiseworthy actions, this has no consequences for OIC. The reason, of course, is that when an agent does the morally right thing for the right reasons, she is thereby doing what she ought to do, and since she is doing it, she clearly can do it. The problem for this principle is related to situations in which, apparently, the agent does something morally wrong, which she ought not to do, and is blameworthy for doing it though she could not have avoided doing it. But we have not yet found any cases of this sort. So the falsity of PAP for what concerns praiseworthy actions, if Luther cases actually showed that, would have much less damaging consequences for moral responsibility, for that falsity would leave the OIC principle intact.

Even so, it does not seem to me that Luther cases actually show the falsity of PAP for morally praiseworthy actions. Some libertarian philosophers, who have become convinced by these cases of the falsity of PAP itself, have reacted to this challenge, following Robert Kane, by withdrawing to other ways of stating the necessity of alternative possibilities for moral responsibility. This has also been a frequent libertarian strategy in the face of Frankfurt cases (remember Van Inwagen's PPP1, for instance) or self-trapping cases (think of Lamb's Weak Principle of alternate possibilities). According to Kane, even if what a person does at a certain time is an unavoidable consequence of her motives and character, so that she could not have done otherwise at that time, she can still be morally responsible for that action provided that she is responsible for presently having those motives and character. And she can be responsible thereof in virtue of earlier choices and actions by means of which she built up such motives and character and regarding which she could have done otherwise. So, in Kane's view,

> …[t]he A[lternative] P[ossibilities] condition can…withstand attacks like Dennett's, but only if we put a gloss on it. Not *all* of our morally responsible choices or actions (those for which we are truly praiseworthy or blameworthy) have to be such that we could have done otherwise with respect to them directly. Yet *some* of the choices or actions in our life histories must satisfy AP if we are to be ultimately morally responsible for anything we do.
>
> (Kane 1996:40)

Kane's response has some similarities to Lamb's reaction to the lifeguard example. Faced with a presumptive case in which someone seems clearly morally responsible for an act with respect to which she could not have done otherwise, both Kane and Lamb go back to a different time and action with respect to which the agent could have done otherwise, so that she bears responsibility for the former, to which she had no alternative, indirectly, through her responsibility for the latter, to which she did have. According to Kane, however, responsibility for an action traces back, ultimately, to responsibility for one's own character and motives, to what Charles Taylor (1982) has called "responsibility for self". Ekstrom follows Kane's steps when she claims that one is responsible and praiseworthy for a certain action, even if at the time one could not have done otherwise, only because "at some points in time, when making decisions that led to his forming the character he has, he really *could* have, in a categorical sense, done otherwise" (Ekstrom 2000:165).

According to this kind of reaction to Luther examples, then, alternative possibilities can ground moral responsibility for a certain action in a very indirect and derived sense: an agent can be morally responsible for a certain action with respect to which she could not, at the time, have done otherwise, but only if she is morally responsible for herself, and she is so only if, in the past, she had alternatives to being the sort of person she now is. One worry that this response to Luther cases may raise, however, is that the responsibility for an action to which the agent has no alternative possibilities is, so to speak, bought on credit. We shall come back to this question, but let us now express the suspicion that the debt might simply be too high. One problem to anticipate is the following. It may seem that Luther's act is performed on the basis of very deep convictions and reasons, and that therefore he is more clearly morally responsible for this

act, to which he claims he has no alternative, than for more stormy and agitated past choices made in the absence of such meditated and strong reasons as he now has. But if his responsibility for his present act derives from responsibility for those former choices, this plus of responsibility looks a bit mysterious. Moreover, Kane's move, namely to shift alternatives back into the agent's self-forming process, seems to assume that at some time an agent can be so completely formed that some actions and decisions just flow inevitably out of her character and motives. It seems to me, however, that agents are never so fully and readily constituted. All this counsels a more direct response to Luther cases.

In connection with these remarks, it is worth noting that, in Kane's view, the alternative possibilities condition is required for moral responsibility also in an indirect sense, namely by being constitutive of a broader necessary condition for moral responsibility, which Kane labels "Ultimate Responsibility". Ultimate Responsibility is Kane's proposal for spelling out what is contained in the intuitive view that moral responsibility for a certain action requires that the action starts or originates, in a rather strong sense, in the agent herself, which in turn requires her enjoying some degree of autonomy or self-determination. We shall call this condition "ultimate control". For Kane, then, alternative possibilities are relevant to moral responsibility through their connection with a metaphysical condition for moral responsibility, namely ultimate control. In this context, alternative possibilities are required for both blameworthiness and praiseworthiness. Kane's, however, is not the only way in which the importance of alternative possibilities for moral responsibility can be viewed or explained. As we have suggested, it is also possible to see this importance in connection with the OIC principle. For someone primarily interested in preserving this principle, it is enough that alternative possibilities are required for blameworthiness, and not necessarily for praiseworthiness as well, for, as long as one does what is morally required, one can do it, and the principle is safe, whether or not one could do otherwise. Someone who adopts this perspective will not see Luther cases as especially worrying. However, these two motivations for defending the alternative possibilities condition are not mutually exclusive, and if someone has both, she will see alternative possibilities as required for moral responsibility in general. She will then be worried about Luther cases, for these may seem to show that praiseworthiness does not require alternative possibilities.

The considerations in the last two paragraphs counsel facing the challenge that Luther cases raise for PAP in a direct way, trying to show that these cases do not actually prove that alternative possibilities are not required for praiseworthy actions.

Let us suggest, to begin with, that the three cases we have considered are relevantly different. As for the first, we have already argued that it is far from convincing. It is hard to think that the woman's phrase, "I couldn't resist", is literally true. Ekstrom says, correctly in my view, that such avowals are usually exaggerations for the sake of modesty or humility. I would add, however, that we also use them in order to discharge another person from an excessive debt of gratitude for something we did for her, that is, in order to suggest that our act did not really have that much merit, that it was not so *praiseworthy*. This is interesting because it reveals an assumption that not having alternatives detracts praiseworthiness or merit from an act. However, if the example itself is less than convincing, we should concede that this response does not sufficiently establish the opposite thesis either.

Luther's example is less easy to deal with. But still we think it is far from decisive. Luther's "I can do no other", unlike the "I couldn't resist" in Wolf's example, is not an expression of modesty or humility, but rather of the strength of his resolution to break off his allegiance to the Roman Church. As Dennett correctly says, Luther was not trying to eschew his responsibility. Quite the contrary seems true: he was thereby assuming full responsibility for his decision. This decision, however, was not a sudden impulse. Rather, "I can do no other" conveys just the opposite: that there were just too many and too deep reasons in favour of his decision to take the possibility of a withdrawal seriously.

Luther's example was certainly not a Buridan case, but this does not mean that he did lack any reasons against his resolution to break off his obedience to the Pope. One can think of just too many of them, including tiredness and strong pressure coming from the Roman Church's representatives. If we think of Luther's act as praiseworthy it is, in part, because we can see how understandable and easy it would have been for him to give in to that pressure, that is, because we see that Luther had alternatives. In fact, the inner, stormy struggle he went through before making up his mind shows that his final decision was not a matter of course for him and that he saw there were reasons, even powerful ones, against it. If he did not do otherwise, it was not because he lacked reasons for doing so.

At this point one may tend to think again that this lack of balance between his reasons deprives Luther of alternatives. Ekstrom shows this tendency when she writes that "being pushed into deciding in a certain way by *anything*—whether one's grandmother, one's genetic blueprint, or *overwhelmingly powerful considerations*—is antithetical to free agency" (Ekstrom 2000:129, my emphasis). But a reasonable incompatibilism will have to concede to compatibilists that there is a difference between being led to act by brute causes and by rational considerations. Viewing *any* factor that can cause a decision or action as equally threatening to one's freedom is the sort of move that libertarians should not be tempted to make if they do not want to face the usual charges of arbitrariness or irrationalism. We have argued in several places that moral responsibility has rationality conditions, so that deciding for reasons, even "overwhelmingly powerful" ones, should not be seen as opposed to freedom, but as constitutive of it as a basis of moral responsibility. Viewing obedience to one's conscience and reasons as opposed to freedom is to use "freedom" to mean something that has little to do with moral responsibility. Now, her conscience and reasons may underlie an agent's avowal that she could do no other; however, she does not mean by it that it is literally, physically impossible for her to do otherwise, but that doing so would be crazy, irrational or immoral. But sometimes, too often in fact, people do crazy, irrational or immoral things. So when they do what they ought to do and do it for the right reasons, they are praiseworthy for that, but part of the explanation is that they could have done otherwise, and that it would be just too explicable if they had done so. So I simply find Dennett's example unconvincing.

The third example is Dennett's claim that it would be impossible to induce him to torture an innocent person by an offer of a thousand dollars (Dennett 1984:133). I imagine that we can all think of lots of awful things that it would be impossible to induce us to do. He seems to present this example as an additional illustration of the lesson he intends to draw from the Luther example. However, these two examples seem to me to be importantly different. One difference is that, unlike what happens in Luther's case, here the rejected alternative can be very close to being literally, even physically

impossible for most of us to choose, in the sense that we may feel simply incapacitated to go for it. Merely imagining doing so may make us feel sick. However, if Luther had withdrawn from his resolution to break off his allegiance to Rome, we might have found this sad and morally wrong, but not inexplicable. As we suggested, there are many times at which human beings act against their moral convictions and values owing to many different reasons. And reasons of this kind were certainly there for Luther. Had he given in to the Roman authorities, this would not have lacked a rational explanation. But I cannot imagine Dennett accepting one thousand dollars for torturing an innocent person. I would not know how to explain that in rational terms. Perhaps we should have recourse to psychopathology. There are, however, hard as this can sound, people who accept offers of this kind, and even less, not just for torturing but also for killing innocent people. And they are not necessarily psychopaths. Rather, what the fact that they accept this sort of offer shows is that their sensitivity to moral motivations is seriously impaired, if not fully absent. They are, so to speak, morally blind. I would tend to think that they may be legally but not morally accountable. So not being able to choose alternatives involving morally awful things is a condition for being a morally responsible agent in the first place, that is, for being an agent to whom moral responsibility can justifiably be ascribed. Now, being sensitive to moral reasons and motivations, and so meeting this minimal condition for being a morally responsible agent, is surely a good thing. However, our intuitions about praiseworthiness may be much less firm here than in a case like Luther's. We admire Luther's moral worth and determination to go ahead, in spite of pressures and all-too-human motives, with what he thought to be deeply important and valuable. But if Dennett were actually offered one thousand dollars for torturing an innocent person and he rejected it, I think that our admiration for him would not significantly increase. Correspondingly, if Luther had chosen to withdraw, he would have been morally blameworthy, and therefore still a morally responsible agent. But if Dennett accepted the bribe, rather than straightforwardly judging him blameworthy, I think we should and would have serious doubts about whether ascribing moral responsibility to him would still make sense from then on.

These considerations show, I hope, that Luther and Dennett's examples are very different. Dennett's bribe example does not show that alternative possibilities are not required for moral responsibility, of either the praiseworthy or the blameworthy varieties. What it rather shows is the minimal level beyond which moral responsibility ascriptions, of either variety, can begin to make sense. And beyond this level, alternative possibilities seem clearly required for blameworthiness, and also for praiseworthiness.

Conclusion

In this chapter, we have argued that PAP is not falsified by any of the most important lines of attack on it. We have defended alternative possibilities as a necessary condition for moral responsibility. Now, if determinism rules out alternative possibilities, in defending their necessity for moral responsibility we have thereby given support to incompatibilism. But incompatibilism is premise B of SMR, the sceptical argument about moral responsibility. So, somewhat paradoxically, in this general framework, our considerations in favour of PAP, certainly akin to libertarianism, none the less reinforce

the prospects of scepticism about moral responsibility, in reinforcing one premise of SMR. Let us now turn to another line of argument in favour of SMR's premise B (incompatibilism). According to this line, a necessary condition for us to truly deserve moral praise or blame for our choices and actions is that we have deep, ultimate control over them. Only then can we truly be said to be their ultimate sources and authors, and only then can it be justified to hold us morally responsible for them. This sort of control, however, seems clearly impossible if determinism is true.

3
Moral responsibility and control
(SMR's premise B)

We have followed a line of argument that, if sound, already establishes premise B of SMR. This line of argument runs as follows: determinism rules out alternative possibilities; alternative possibilities are necessary for moral responsibility; therefore determinism rules out moral responsibility. This is an indirect argument for incompatibilism, because it leads to the conclusion that moral responsibility and determinism are incompatible through a detour consisting of two premises about alternative possibilities.

We have argued in favour of these two premises. But many philosophers, both compatibilists and incompatibilists, are not convinced of their truth. The premise about the necessity of alternative possibilities for moral responsibility has come to be seen as particularly contentious, especially through the influence of Frankfurt cases. Rejection of this premise (or of the premise that determinism excludes alternative possibilities) leads to rejecting the soundness of this indirect argument for incompatibilism. But incompatibilism (SMR's premise B) can also be defended by a second line of argument. The argument runs roughly as follows: 1) If an agent is to be morally responsible for a certain decision or action of hers, she has to have ultimate control over this decision or action, so that she can be said to be their ultimate source. 2) Determinism is incompatible with an agent's having ultimate control over her decisions and actions, and so with her being their ultimate source. 3) Therefore, if determinism is true, moral responsibility is not possible (SMR's premise B). Let us comment on the premises of this second argument for incompatibilism.

Ultimate control and determinism

The requirement of ultimate control derives quite naturally from reflection on the nature of moral responsibility ascriptions. Our notion of responsibility, whether moral or otherwise, is, at least partially, a causal notion: to judge that someone is responsible for A is, among other things, to judge that she is, at least partially, a cause or origin of A. Remember that the ancient Greeks had one single word, namely *aitia,* to denote both cause and blame, and we still sometimes use "responsible" in a purely causal sense, as when we say that a stroke of lightning was responsible for the fire in the wood. However, when we say of an agent that she is morally responsible for A, we are not only saying that she was a cause or origin of A. Attributions of moral responsibility, of moral praise- or blameworthiness, are usually strongly evaluative, and it is not only the action, omission or consequence of either that is judged good or bad, praise- or blameworthy, but also, and

especially, *the agent* herself, whose worth and value are thereby potentially increased or diminished. If moral responsibility ascriptions are to be justified, the agent truly has to deserve them, and true desert requires her having genuine, ultimate control over that for which she is held morally responsible. The requirement of ultimate control or self-determination corresponds to this depth and personal significance of moral responsibility ascriptions. If the agent did not control the process that led to her action or omission, if, in this process, she was driven by factors beyond her reach, as happens for instance in cases of strong coercion, hypnosis or grave pathological compulsion, one is not justified in praising or blaming *her*. She does not truly deserve praise or blame. It seems, then, that the grounds on which moral responsibility ascriptions are made have to be appropriately in the agent's hands. Otherwise, she can find herself blamed or praised, and so diminished or enhanced as a person, owing to factors fully beyond the limits of her influence. The requirement of ultimacy, then, is closely connected with our idea of free-willed agents as initiators of changes in the world, quite unlike ordinary instrumental or intermediate causal factors; it lies at the very root of the significance we ascribe to having a free will. This is the idea that Robert Nozick (1981:311–12) expressed by saying that to have a free will is to have "originative value", that is, the power to create and introduce new value in the world, instead of being a mere carrier of what is already, implicitly, there. This aspect of free will is inherited by, and constitutive of, our concept of a morally responsible agent.

But what is included in this notion of ultimate control or true self-determination? A general characterization can be found in Susan Wolf, who clearly sees the attraction and importance of this requirement, even if she ends up rejecting it as impossible to attain. According to her, "it seems that…there is a requirement [for moral responsibility] that the agent's control be ultimate—her will must be determined by her self, and her self must not, in turn, be determined by anything external to itself" (Wolf 1990:10). Following Kant, Wolf calls this requirement of ultimate control "autonomy". And she clearly sees its connection with the significance of free will and moral responsibility that we have just talked about: "It makes sense that beings who can purposefully initiate change…should have a special significance, for they are sources of value (and disvalue) rather than mere carriers of it" (Wolf 1990:10).

A more detailed characterization of this requirement for moral responsibility can be found in Robert Kane, who dubs it "Ultimate Responsibility". According to Kane, if an agent is to be ultimately responsible for (in ultimate control of) an action of hers, she has to be personally responsible not only for the action itself but also for the causes or sufficient grounds of that action. Moral responsibility requires controlling our actions, so to speak, all the way down. Let us quote Kane:

(UR) An agent is *ultimately responsible* for some (event or state) E's occurring only if (R) the agent is personally responsible for E's occurring in a sense which entails that something the agent voluntarily (or willingly) did or omitted, and for which the agent could have voluntarily done otherwise, either was, or causally contributed to, E's

occurrence and made a difference to whether or not E occurred; and (U) for every X and Y (where X and Y represent occurrences of events and/or states), if the agent is personally responsible for X, and if Y is an *arche* (or sufficient ground or cause or explanation) for X, then the agent must also be personally responsible for Y.

(Kane 1996:35)

So, if, at a certain moment, a decision or an action of ours is explained by our motives and character, we must be personally (as opposed to collectively) responsible for the latter in order for us to be ultimately responsible for the former. Especially important in this regard are, in Kane's view of the ultimate control requirement, what he calls "self-forming willings", undetermined, regress-stopping choices or acts of will by means of which an agent forms her own character. These constitute the absolute, ultimate origin of an agent's actions and so are meant to ground ultimate control and to justify attributions of moral responsibility.

Though Kane conceives the ultimate control or self-determination requirement as the central condition for moral responsibility, he considers this condition as itself dependent on the alternative possibilities condition: unless there are some actions with respect to which an agent could, categorically, have done otherwise, she could not have ultimate responsibility for any action of hers, nor could she, therefore, be morally responsible for it. On this assumption, we might have a unified line of argument in favour of SMR's premise B. The argument would run thus: 1) Ultimate control over one's decisions or actions is a necessary condition of moral responsibility for those decisions and actions. 2) Ultimate control requires alternative possibilities of decision and action. 3) Determinism rules out alternative possibilities of decision and action. 4) So determinism rules out ultimate control. 5) Therefore determinism rules out moral responsibility.

But incompatibilists can also base their case exclusively on the ultimate control condition, with no appeal to alternative possibilities. They can be impressed by Frankfurt or other sorts of cases and be led thereby to reject the alternative possibilities condition for moral responsibility while insisting on the necessity of ultimate control or self-determination. They can further contend that determinism rules out the satisfaction of this latter condition and therefore rules out moral responsibility. Their line of argument would include only steps 1, 4 and 5 of the unified incompatibilist argument that we have just set out, and would be equivalent to the second argument for incompatibilism that we presented at the beginning of this chapter, which deals only with ultimate control, not with alternative possibilities. Derk Pereboom, for example, would endorse this position, though he further holds that (ultimate) control is excluded by (certain forms of) indeterminism as well. For this reason, he defines himself as a hard incompatibilist, and not only as a hard determinist.

There is little doubt that satisfaction of the ultimate control condition is incompatible with determinism, that is, that premise 2 of the incompatibilist argument at the beginning of this chapter is true. Determinism is the thesis that the conjunction of a full description of the past and the natural laws logically implies all truths. Now, with the possible exception of an absolute or ultimate origin, no event, state of affairs or configuration can be an ultimate, initiating cause if that thesis is true. For any event, including choices and actions, and for any configuration, as persons presumably are, there is a complete explanation in terms of the past and the natural laws. So, with that possible exception,

nothing in a deterministic universe can be an ultimate cause or origin; everything that is part of that universe is itself the product of prior causal factors, and so at most an intermediate cause. But the possibility of ultimate control over one's actions or the consequences thereof requires the possibility of ultimate causes: one can have ultimate control over one's own actions only if these can have an ultimate, causally undetermined cause. So determinism rules out the possibility of ultimate control.

If this is so, then compatibilists, who want moral responsibility to be consistent with determinism, will have to deny premise 1 of our second incompatibilist argument, that is, they will have to deny that ultimate control or self-determination is a necessary condition for moral responsibility. This denial does not commit compatibilists to reject any control or self-determination condition on moral responsibility. They certainly, and typically, agree that certain forms of lack of control over one's actions, such as coercion or strong pathological compulsion, rule out moral responsibility, so that agential control is required for moral responsibility. But they strongly object to the idea that the required control has to be ultimate and, therefore, incompatible with determinism. Compatibilists typically view the incompatibilist requirement of ultimate control as conflicting with any reasonable naturalistic view of human beings. Whether our world is deterministic or not, we human beings are contingent and passing configurations formed entirely out of physical and biological factors which were already there before we were born and over which, and the way in which they have been structured, we have had no control whatsoever. There is also reason to think that at least some of our basic psychological traits come with this physical and biological heredity. Besides, we do not have any control over the social and physical environment in which we spend the first years of our lives, and this environment shapes our selves in a decisive way and determines our opportunities for education and further personal development. How, then, could a human being become a centre of ultimate control and self-determination and be an ultimate source of her actions if she herself is constituted out of an array of factors beyond her control? In the light of naturalism, ultimate control looks like an unreasonable demand, which fosters scepticism about moral responsibility.

The challenge for the compatibilist is to put forward a concept of control or self-determination that is both consistent with naturalism and determinism and robust enough to ground moral responsibility ascriptions. She has to show that, even if our world is deterministic, an agent can enjoy a degree of self-determination and control over some of her actions that can justify ascriptions of moral responsibility for them. The central tenet of compatibilism is to look, within the causal network that constitutes a human being, for the sort of causal factors that can justify talking about an agent's self-determination and control over her own actions. It is vain to try to escape from the causal chain that has given rise to us as human beings, for this would mean to try to escape from ourselves and to lose precisely what we are looking for, namely the source of our own actions and decisions. The complex configuration that is a human being already contains the sort of inner structures that generate free and responsible actions. Acting in a self-determined way, so that the agent can be rightly held to be the origin of that act, does not require causal gaps, but rather the right sort of causal history. In searching for an absolute, causally undetermined origin, incompatibilists undermine the very foundations on which moral responsibility can plausibly rest, for in this foolish search they overlook the factors and distinctions that can make sense of the notions of control and self-determination.

There is, for example, a big difference between acting on the basis of one's own deliberation and decision and acting because of such factors as coercion or compulsion. This sort of difference may be all that is needed to distinguish between being, and not being, in control of one's actions. To hold that, if both processes are causally determined, there is no relevant difference between them for what regards their consequences on moral responsibility, is to lose sight of our ordinary practice of moral responsibility ascriptions and to look instead for an unattainable myth.

Let us now look at some compatibilist attempts to propose non-ultimate (non-autonomous, in Wolf's terms) forms of control or self-determination that could ground moral responsibility. Our main task will be to see whether compatibilists are right in claiming that ultimate control or self-determination is not required for moral responsibility.

Classical compatibilism: actions, desires and the self

Under the label "classical compatibilism", which we borrow from Gary Watson (cf. Watson 1987:145), we include the views of such thinkers as Hobbes, Hume and Ayer. Their position is stronger than mere compatibilism. Besides holding that determinism is compatible with moral responsibility and with the freedom relevant to it, they also believe that determinism is true. Some of them may also believe that it is required for freedom and moral responsibility. Though terms such as "control" or "self-determination" are not theirs, the freedom they speak about is a certain construal of the concept expressed by those terms. Freedom to do something is a certain kind of ability or power, namely the ability or power to do what one wants and decides to do, and to do it because one wants and decides to do it. It is, then, a causal power. And to act freely is to exercise that power, namely actually to do what one wants and decides to do because one wants and decides to do it. In this context, freedom to do otherwise is to have that ability with respect to an alternative action: we are free to do otherwise just in case we could or would do otherwise if we wanted and decided to. The conditional analysis of the alternative possibilities condition is clearly suggested by the classical compatibilist notion of freedom.

Freedom, then, is compatible with causal determinism in that an agent can enjoy the power which freedom consists in even if her desires and decisions have deterministic causes. What is required for freedom is to have desires and to be able to decide and act upon them. It is not required that these desires and decisions be causally undetermined. On the other hand, determinism may even be required for freedom precisely because freedom requires one's desires and decisions to be able to cause one's actions. An agent's act is not free, then, when this act is caused not by her own desires and decisions, but by other, external causes. The distinction between free and unfree actions is, then, internal to the causal deterministic network. Freedom, according to Hobbes and Hume, is not opposed to causation, but to coercion or constraint. One's desires and decisions cause one's actions: they do not constrain them.

We can now see how this notion of freedom is intended to make sense of control and self-determination and, so far, of moral responsibility as well. Self-determination, on this view, is determination by the self, with no requirement that the self be undetermined. And, in Hobbes's and Hume's views, one's self is a psychological structure that is centrally constituted by one's motives and desires. So, when a subject acts because of her own motives and desires, it is her own self that is determining her actions, and this is precisely what self-determination amounts to. Moreover, acting in this self-determined way is precisely being in control of one's actions. Again, self-determination and control are not opposed to causation of one's actions, but only to causation of them by factors other than the agent's self, such as external force or constraint. Self-determination thus requires having a self in the above sense: it does not require this self to be causally undetermined.

Freedom, on this view, applies essentially to those events that can be caused by an agent's will, which, in Humean terms, comprises her desires and decisions; paradigmatically, therefore, it is actions that can be free. Acting freely, and being in control of one's actions, is, roughly, doing what one wants and decides to do, or acting on one's own desires and decisions. But, since it is the will that confers freedom on actions by causing them, it does not make sense to speak about the will itself being free. Such expressions as "free will" or "freedom of the will" are absurd. Schopenhauer, whose conception of freedom is very close to classical compatibilism, expressed this point by saying that a man can do what he wants, but not want what he wants. Another way of expressing this idea is to say that someone can act as she wants to act, but not want as she wants to want. Actions can be subject to the will and thereby be free. The will, however, cannot sensibly be said to be subject to itself, nor can it sensibly be said to be free.

Hobbes and Hume typically think of external constraint and coercion when they think of causes that deprive an agent of freedom in that they are external to this agent's self. An agent's desires and motives are, for them, internal to the agent's self and, therefore, do not constrain or compel her actions. But these notions of freedom and of the self, and the corresponding notions of self-determination and control, are too permissive. Desires and motives that are her own can sometimes constrain an agent if, for example, they cause her actions in a compulsive or obsessive, pathological way. In these circumstances, they cause the agent's decisions in a way similar to the way in which external constraints do: in both cases, the agent experiences herself as governed by alien forces. This is an experience of other-determination and lack of control that Hobbes's and Hume's views cannot account for. And this sort of influence on an agent's behaviour is generally seen as diminishing her freedom and moral responsibility, and even, beyond certain limits, as depriving her of them. It was Ayer who took this point into account, restricting the scope of an agent's self by excluding from it desires and motives of a compulsive, pathological sort and assimilating them to external coercion for what regards their effect on freedom:

If I suffered from a compulsion neurosis, so that I got up and walked across the room, whether I wanted to or not, or if I did so because someone else compelled me, then

I should not be acting freely. But if I do it now, I shall be acting freely, just because these conditions do not obtain; and the fact that my action may nevertheless have a cause is, from this point of view, irrelevant. For it is not when my action has a cause at all, but only when it has a special sort of cause, that it is reckoned not to be free.

(Ayer 1954:21)

Ayer does not provide a reasoned principle to distinguish between those mental states that constitute the agent's self and those that are external to it for what concerns their effects on her freedom and moral responsibility. His position, however, seems to be that a free action, one for which the agent can justifiably be held responsible, is one which is caused by the agent's reasons, by those belief—desire sets that, by causing, and causally explaining, her behaviour, make it into a case of rational action. Ayer's view of free action is, then, quite close to Davidson's, which in turn is closely dependent on his own causal conception of intentional action. Davidson's view of intentional and free action can be considered as a refined version of classical compatibilism.

Classical compatibilism gives content to the concepts of self-determination and control while rejecting the requirement of ultimacy. According to this position, an agent can be said to control her actions, and so to be their origin or cause, to a reasonably high degree to justify attributions of moral responsibility. Incompatibilists will find this view of control too weak and insufficient to ground moral responsibility and justify ascriptions thereof: in the absence of ultimate control and self-determination, an agent cannot truly deserve moral praise or blame. Compatibilists will typically reply that ideas of radical control or self-determination, or of an absolute or ultimate origin, are Utopian and, perhaps, even incoherent. They will agree that the control and self-determination they can offer is not ultimate, but will insist that this is all that can reasonably be expected and that wanting more than that is likely to lead to the sceptical conclusion that moral responsibility simply cannot be had by anyone.

However, the classical compatibilist construal of the control condition falls arguably short of accommodating not just incompatibilist intuitions but some important aspects of our ordinary concepts of freedom and moral responsibility as well. On the one hand, the concepts of the self and of the will it makes use of are too slender and poor to make sense of our distinction between those beings to which we naturally ascribe free will and moral responsibility and those with regard to which these ascriptions seem clearly inappropriate. If enjoying self-determination and control in acting, in the sense required for moral responsibility, just means doing what one wants to do, then this condition is satisfied by small children and some higher animals that are plausibly seen to have desires. The construal of this condition clearly seems too weak and permissive to be acceptable. It clearly falls short of providing a strong enough account of our control and self-determination requirement for moral responsibility. It gives no explanation of our refusal to consider small children and higher animals as being sufficiently self-ruled and in control of their acts to count as plausible candidates for the title of morally responsible agents.

On the other hand, even if we leave aside pathologically compulsive desires, not all ordinary desires on which we act are on a par for what concerns our freedom and self-determination. For example, some personality traits that, below a certain threshold, are not pathological, give rise to characteristic desires and motivational states in a subject. Think of envy. There are envious people, just as there are generous, mean and intelligent people. An envious person often desires other people's failure and, sometimes, acts on such a desire. Suppose, none the less, that she is not happy with this feature of her character nor is she when she feels these characteristic desires. They are not pathologically compulsive, but the subject is, so to speak, ashamed of them. She regrets having these desires, as well as her tendency to act on them. And if she acts on these desires, she experiences them as forces somehow alien to herself; she does not feel herself fully self-determined in so acting, or fully in control of her action. But, according to classical compatibilism, this subject's acting on these desires is an example of self-determination and control. Classical compatibilism, then, cannot explain this characteristic experience of other-determination and lack of freedom. Something is wrong with the classical compatibilist construal of the control condition.

These shortcomings of classical compatibilism have given rise to new attempts to account, on broadly compatibilist lines, for the control condition for moral responsibility. In a series of papers, Harry Frankfurt has developed a highly influential approach to this problem. The basis of this approach is his conception of free will.

Frankfurt's view of freedom of the will

According to Frankfurt, classical compatibilism's view of freedom as an ability to do what one wants to do may capture the notion of freedom of action, but not the notion of freedom of the will or of free will, which has been the main concern of traditional philosophical reflection. Correspondingly, the associated notion of control or self-determination also falls short of providing a correct analysis of this freedom-related requirement of moral responsibility. At the root of these shortcomings there is a wrong, over-simplified view of the self. Classical compatibilism has largely underestimated the psychological complexity required for talk of free will to make sense with respect to an agent. As we suggested, agents such as small children and higher animals can enjoy freedom as defined by classical compatibilism, and can exercise this freedom by acting freely. A dog, for example, can have freedom of action: it can sometimes (when, e.g., it is not tied) do what it wants to do and actually do it. But it can hardly be said to have a free will or to be morally responsible for what it does, and, *mutatis mutandis,* the same applies to very young children. The sort of control they have over their actions is not that which we assume in attributions of free will and moral responsibility. What is it that distinguishes them from those agents to whom free will can be meaningfully attributed? People and some animals have desires, and they sometimes act upon them. Frankfurt calls them "first-order desires". First-order desires are desires whose intentional objects are actions: they are desires to do certain things. The will, in Frankfurt's definition of this term, is constituted by those first-order desires that effectively move an agent to act. In this sense, young children and higher animals also have a will. What they lack, however, is reflective conative attitudes: desires about which (first-order) desires they want to have

and be moved by. These reflective desires, whose intentional objects are first-order desires, are called by Frankfurt "second-order desires". Some of these second-order desires, however, are not merely desires to have or not to have a certain first-order desire, but desires that a certain first-order desire be one's will, that is, be effective in moving one to act. These reflective desires, whose intentional object is the will itself, are, in Frankfurt's terminology, "second-order volitions".

The concept of second-order volitions is central in Frankfurt's conception of free will. He illustrates this concept with the case of an unwilling drug addict (cf. Frankfurt 1971:17–18). This unwilling addict, this addict in spite of herself, as we could describe her, has contrary first-order desires. She simultaneously wants to take the drug and to abstain from taking it. But she is not indifferent to these desires in conflict. She has a reflective attitude towards them: it is the latter, rather than the former, that she wants to be her will. It is the latter desire, instead of the former, that she wants to be efficacious in leading her to act. This reflective attitude is what Frankfurt calls a "second-order volition". This unwilling addict is a "person", again in Frankfurt's terms, in that she has second-order volitions: she cares about her will. And this is what is needed for talk about an agent's free will to make sense. Frankfurt construes the concept of free will in analogy to the concept of freedom of action:

> [F]reedom of action is (roughly at least) the freedom to do what one wants to do. Analogously, then, the statement that a person enjoys freedom of will means (also roughly) that he is free to want what he wants to want. More precisely, it means that he is free to will what he wants to will, or to have the will he wants.
>
> (Frankfurt 1971:20–1)

Exercising freedom of the will, then, consists in achieving the harmony between one's will, or effective first-order desires, and one's second-order volitions. And lack of this freedom is experienced as an inconsistency between those two conative attitudes. According to this view, the unwilling addict's will, provided that she finally takes the drug, is not free for what regards this action, for this is not the will she wants to have. Her will, however, can be free with respect to other actions. She is a person, in that she has second-order volitions, and talk about her free will makes sense: she may actually have the will she wants to have, because she *has* desires (second-order volitions) about her will (effective first-order desires). Dogs, however, or even human beings who do not have second-order volitions, not only have no free will: they cannot have it, either. They lack the required structural traits (reflective, second-order conative attitudes) for questions about their free will to be sensibly asked. In Frankfurt's terms, an agent with no second-order volitions is not a person, but a *wanton*. Wantons act on the sole basis of their first-order desires. They can have conflicting first-order desires, but they do not have a reflective desire concerning the desires in conflict. The conflict is solved beyond the agent, so to speak, as a function of the respective strength of those first-order desires. Small children and higher non-human animals belong to this class. But, as Frankfurt ironically points out, some adult human beings can also belong to it. At any rate, all of us behave in a wanton-like way on many occasions, in acting on desires about which we do not have reflective, second-order volitions.

Frankfurt's approach to the free will problem represents a significant advance with respect to classical compatibilism. According to Frankfurt, freedom, as characterized by classical compatibilism, corresponds only to freedom of action. We can also say it is the sort of freedom that a wanton can enjoy. Something similar can be said about the classical compatibilist concept of control or self-determination: it is the sort of control or self-determination that wantons can enjoy, which concerns actions exclusively, not the will. In connection with this, Frankfurt has shown that, contrary to what Hobbes and Schopenhauer contended, the concept of freedom of the will, as these thinkers formulated it, is not absurd: it makes sense to say of an agent that she wants what she wants to want.

Frankfurt's concept of free will is consistent with a broadly naturalistic perspective. On his view, free will, the ability to have the will one wants to have, does not require anything beyond the reach of a scientific outlook: it is fundamentally a matter of having a psychological structure with a certain level of complexity. The structure has to include, at least, reflective second-order conative attitudes. It is not (first-order) desires, as classical compatibilism held, but second-order volitions that essentially constitute the self, and this is why some of our first-order desires can be experienced as somehow external to us. This is what happens, for example, in the case of a person who wants to give up smoking without succeeding. The unwanted, though dominant, desire to smoke, Frankfurt remarks, is "*external* to the volitional complex with which the person identifies…" (Frankfurt 1987:165). As in classical compatibilism, self-determination is understood as determination by the self, but it is by appeal to these reflective attitudes, not to first-order desires, that the notion of control is to be explicated in connection with free will. To have a free will is to be able to have volitional control over one's will, in the sense of having the will one wants to have. And to exercise this ability is to control one's will effectively: to have the will one wants to have, because one wants to have it. This is also what it means, for an agent who has a free will, to be appropriately in control of her actions, beyond the weak sense in which wantons can also control them. It is a matter of an agent's acting on desires with which she identifies and which she endorses as motives of her actions. Control or self-determination is, then, a causal power: the power of one's self, or second-order volitions, to bring about one's actions. In some sense, Frankfurt's approach to freedom and self-determination can be said to reproduce the basic structure of classical compatibilism, only adding one further level to it.

Frankfurt's perspective has the resources to solve some of the problems classical compatibilism stumbles on. It is pretty obvious how it can account for the difference between animals and adult human beings for what concerns freedom of the will and moral responsibility. And it can also explain the experience of heteronomy that accompanies some cases of acting on one's own desires, as happens with our example of an envious person who hates her envy. This experience can be explained by the discrepancy between first-order and second-order desires or, more exactly, between the agent's will and her second-order volitions.

Frankfurt's view of free will departs considerably from the libertarian tradition, but it connects with it in assuming that free will requires alternative possibilities:

A person's will is free only if he is free to have the will he wants. This means that, with regard to any of his first-order desires, he is free either to make that desire his will or to make some other first-order desire his will instead. Whatever his will, then, the will of the person whose will is free could have been otherwise; he could have done otherwise than to constitute his will as he did.

(Frankfurt 1971:24)

So the sort of control involved in having a free will is "dual control" (Kane) or "regulative control" (Fischer), control over which alternative possibility will become actual: over whether to have a certain will (an effective first-order desire) or an alternative will instead. However, we know that Frankfurt denies that alternative possibilities (dual or regulative control) are required for moral responsibility. For a person to be morally responsible for something she has done it is not required that she could have done otherwise. This is no news for us. But it is also not required that she could have had a different will, that is, that she could have made a different desire move her to act than that which actually moved her. So, according to Frankfurt, the question of how the locution "could have done otherwise" should be understood is important to the theory of freedom, but it is irrelevant to the theory of moral responsibility. For a person can be morally responsible for something she did even if she was not "in a position to have whatever will [s]he wanted" (Frankfurt 1971:24).

The rejection of the alternative possibilities condition for moral responsibility, but not for free will, leads Frankfurt to deny that free will is itself necessary for moral responsibility. This is an important departure from tradition, both incompatibilist and compatibilist, and it leads Frankfurt to advocate a divorce between the theory of free will and the theory of moral responsibility. An agent can be morally responsible for having done something even if, in the preceding sense, her will was not free at all, so that she could not have acted on a different first-order desire than the one she actually acted on. In our opinion, there is no reason for this strange divorce. Frankfurt's position would gain power and coherence if, consistently with his rejection of the alternative possibilities condition for moral responsibility, he rejected this condition for free will as well. If free will is an agent's ability to have the will she wants to have, it seems that she exercises this ability, and so enjoys free will, if she actually has the will she wants, by endorsing and identifying with the first-order desire on which she acts. There does not seem to be a necessary condition for exercising this ability that she could have wanted and actually had a different will instead, or that she could have identified with a different first-order desire. Of course, I do not think myself that this is a correct position. But the fact that, if I am right, Frankfurt's concept of free will can do without alternative possibilities strongly suggests that it does not correspond to our ordinary concept. Being able to act on a desire one wants to act on does not seem to be what we understand by having a free will. We shall also argue that Frankfurt's conception of moral responsibility does not correspond to our ordinary concept of it either.

Frankfurt's view of moral responsibility

Though, according to Frankfurt, as we have seen, free will is not required for moral responsibility, he none the less holds that a freedom-related condition is actually required for it. According to Frankfurt, the assumption that a person is morally responsible…

…*does* entail that the person did what he did freely, or that he did it of his own free will. It is a mistake, however, to believe that someone acts freely only when he is free to do whatever he wants or that he acts of his own free will only if his will is free. Suppose that a person has done what he wanted to do, that he did it because he wanted to do it, and that the will by which he was moved when he did it was his will because it was the will he wanted. Then he did it freely and of his own free will.

(Frankfurt 1971:24)

What is required for moral responsibility, then, is acting freely, or of one's own free will. This is what corresponds, in Frankfurt's theory, to the control or self-determination requirement for moral responsibility. Unlike the control related to free will, it is a non-dual, non-regulative control; it is an agent's control over the actual sequence that leads to her action, not necessarily over an alternative sequence. Acting freely or of one's own free will, as can be seen in the preceding quotation, requires the same level of internal complexity as having a free will, namely a psychological structure that includes second-order volitions. It also involves the agent's endorsement of her will, the first-order desire on which she acts, by means of a second-order volition. But, unlike free will, it does not involve the possibility of endorsing a different will. This is one queer consequence of Frankfurt's distinction between free will and moral responsibility for what regards the alternative possibilities condition. Let us now examine Frankfurt's view of moral responsibility in more detail.

Frankfurt's theory of moral responsibility can plausibly be seen as an answer to the question of what is involved in the condition of control or self-determination that is relevant to moral responsibility ascriptions. Against classical compatibilism, he clearly holds that acting on a desire one happens to have is not sufficient for having the relevant control over one's action. The desire on which one acts also has to be under the agent's control, and this means that this desire has to be one that the agent wants to have and to move her to act. This wanting is a second-order volition. Now, an agent has the relevant sort of control over her desires in so far as she identifies with them. As for first-order desires, she identifies with them by having a second-order volition. Concerning second-order volitions, the agent identifies with them simply in that they *constitute* her self. Unlike her first-order desires, an agent cannot take the attitude of a passive spectator towards her second-order volitions and view them as forces alien to her self. So, in acting on them she acts in a self-determined fashion. It is worth pointing out, however, that, in later writings, Frankfurt revises this position and acknowledges that a person may be no less wanton with regard to her second-order volitions than "a wholly unreflective creature is with respect to its first-order desires" (Frankfurt 1987:165). To deal with this problem, then, he insists on the importance of an agent's reflective and wholehearted identification with her second-order attitudes. Anyway, it is by means of this two-level identification that the agent performs her actions with the relevant sort of control and self-

determination, thereby fully satisfying this condition for moral responsibility. What Frankfurt calls "acting freely" or "acting of one's own free will" corresponds to this condition.

Frankfurt's theory of control or self-determination is, so to speak, a purely *structural* or *formal* theory. What is involved in satisfying this condition is just a psychological structure of a certain complexity. There are no requirements about the content of the relevant attitudes, first- or second-order, that constitute the structure. It can also be said to be an *ahistorical* theory. According to him, the causal origin or history of that psychological structure is irrelevant to the agent's satisfying the control or self-determination condition. It is by reflectively identifying herself with the springs of her actions that a person can be said to perform those actions freely and to be morally responsible for them. The way in which the reflective identifications were causally produced and came into existence is fully irrelevant to these matters (cf. Frankfurt 1975:54). In later writings, he criticizes Aristotle, in the same vein, "because of his preoccupation with causal origins and causal responsibility" (Frankfurt 1987:171). Finally, it is an *actual sequence* theory, in that, as we have seen, control or self-determination (acting freely or of one's own free will) does not require the availability of alternative pathways in which an agent identifies with a different will or first-order desire. What is relevant to control or self-determination, for what concerns moral responsibility, is only the constituents and structure of the actual causal process that leads to the action, not whether there are alternative processes or what happens in them. Now, since Frankfurt rejects the alternative possibilities condition for moral responsibility, and control or self-determination is, for him, the only freedom-relevant condition of moral responsibility, his theory of moral responsibility can also be said to have those three features.

Self-determination or control is generally held to be necessary (and, on some accounts, sufficient) for moral responsibility. Frankfurt certainly thinks that this condition, on his own interpretation of it, is necessary. In his 1975 paper, he distinguishes situations in which "what motivates [the agent's] action is a desire by which…he does not want to be moved to act" (Frankfurt 1975:48). These are, in Frankfurt's terms, type B situations. And he also distinguishes type C situations, in which "the agent acts because of the irresistibility of a desire without attempting to prevent that desire from determining his action" (Frankfurt 1975:49), as happens in cases of acting under a strong threat. He further holds that in these two kinds of situations, B and C, where no second-order volition backs the agent's first-order desire, she is not morally responsible for what she does. In those situations, the agent does not "act freely", in Frankfurt's sense, and therefore she is not morally responsible for what she does.

Moreover, Frankfurt seems to think that, provided that other, cognitive conditions are fulfilled, self-determination, as he interprets it, is also sufficient for moral responsibility. This is what he seems to hold when he writes that "to the extent that a person identifies himself with the springs of his actions, he takes responsibility for those actions and acquires moral responsibility for them" (Frankfurt 1975:54).

Now, though some sort of control or self-determination is, by virtually anybody's lights, at least necessary for moral responsibility, it is doubtful that what Frankfurt offers as a construal of this condition is, in fact, either necessary or sufficient for it. If so, there is reason to doubt that this construal corresponds to the concept of self-determination or control that is in play in the context of ordinary attributions of moral responsibility.

For what concerns the necessity claim, we can use the example, mentioned on page 94, of an unwilling envious person. Suppose that this person is Harry, a student who actually possesses a book that is highly relevant for a forthcoming examination. Peter, a brilliant colleague of his, whom he envies, asks him whether he has this book, in order to have a look at it. Harry, who strongly desires Peter's failure in the examination, tells Peter that he does not have the book. In acting on this desire, derived from his envy, Harry is not proud of himself. He hates his envy and in fact has a second-order volition not to act on this desire; he does not identify with it, but the temptation is too strong and he finally just acts on it. Assuming that this desire is not compulsive, I think it is fair to judge that Harry is morally responsible for not lending the book to Peter, even if no second-order volition backs his will. It seems, then, that Frankfurt's construal of the self-determination condition is not necessary for moral responsibility. Harry does not act freely, in Frankfurt's terms, but he is morally responsible for his action.

Our judgement presumably changes if we think of Frankfurt's own example of an unwilling drug addict. If the addict's craving for the drug is overwhelmingly strong and finally overcomes his second-order volition that this desire not be her will, we may think that she is not morally responsible for her action (leaving aside the question of how she became an addict). She, like Harry, does not act freely, in Frankfurt's terms. Now, although Frankfurt's view can explain our judgement of this case, it cannot explain our judgement about Harry, nor can it account for the difference between these two judgements. This suggests that a Frankfurtian explanation is not correct concerning the unwilling addict either. There is, however, a straightforward and very plausible explanation of the difference, namely that Harry, unlike the drug addict, had dual or regulative control over his action: he could have done otherwise. Harry ought, and could, have lent the book to Peter, and that is why he is morally responsible for not having done so. But, of course, Frankfurt cannot avail himself of this explanation.

The claim about the sufficiency of Frankfurt's construal of control or self-determination can be challenged by appealing to what Gary Watson (Watson 1987:151) has called "Brave New World cases". Brave New World citizens can satisfy Frankfurt's conditions for moral responsibility. Even if their second-order volitions arise out of a process of systematic conditioning or brainwashing, they none the less have those attitudes. They thus enjoy the required psychological complexity to act freely and be morally responsible for their actions. Remember that Frankfurt's is an ahistorical theory of control and moral responsibility, so that the origin of an agent's second-order volitions and of her identifications with them and with her will is not relevant for her moral responsibility. Even if their second-order volitions have their origin in a planned process of conditioning, Brave New World citizens can still take responsibility for the springs of their actions and even identify with them decisively and reflectively. But consider that their taking responsibility for the springs of their actions can itself be the result of conditioning: they can be conditioned to take responsibility for these springs and to identify with them. To bite the bullet and insist that, in these circumstances, they are

morally responsible agents simply runs too wildly against common intuitions about these cases. Consider that, in ordinary circumstances, a person can be led to take responsibility for something she is not actually responsible for. Suppose, for example, that a person is caused to believe, through a conspiracy of her relatives and acquaintances, that she has committed a dreadful act which in fact she has not committed. It may even seem to her that she "remembers" having performed the act. Now, even if she finally takes responsibility for this act, she is not responsible for it. To hold that she becomes responsible for that act by taking responsibility for it is not only false but also a deep offence against our sense of justice. So, in Brave New World contexts, the fact that citizens take responsibility for the springs of their actions by identifying with them cannot be sufficient for them to be actually responsible for those springs and for the resulting actions. And this is so no matter how "decisive" (Frankfurt 1971) or "wholehearted" (Frankfurt 1987) their identifications may be.

But Brave New World cases, seen from a Frankfurtian perspective on control, also give rise to what seems an unacceptable paradox. It is plausible to suppose that, if the conditioning and brainwashing process is carefully and deeply applied, citizens will only act on desires which are backed by second-order volitions and so will always identify with their wills. This means not only that these people can enjoy Frankfurtian control and self-determination, but also that they could even become *paradigmatic examples* of self-determined agents. And this consequence is really hard to swallow.[1]

What Brave New World citizens lack, it seems, is control over their identifications with first- and second-order volitions. The sort of control Frankfurt offers is not deep enough to sustain moral responsibility. These problems clearly point to the incompatibilist demand of ultimate control as a requirement for moral responsibility: agents have to control the sources of their actions if they are to be morally responsible for them. It seems to be a lack of this ultimate control which underlies our intuitive judgement that Brave New World citizens are not rightly held to be morally responsible for their deeds.

An additional problem for Frankfurt's view of self-determination is related to its purely structural or formal character. The content of the desires relevant to an agent's self-determination is not taken into account. But (and this is something that Frankfurt himself is led to acknowledge) there is no principled reason to assume that the fact that a certain desire or volition is second-order gives it a special role in grounding an agent's self-determination and control over her actions. It is not difficult to think of cases in which it is precisely second-order attitudes that stand in the way of an agent's freedom and autonomy. Certain oppressive, religious or social contexts may lead agents to form second-order conative attitudes opposed to most of their first-order, spontaneous desires for food, pleasure, friendship or love. In these cases, we clearly see these agents' first-order, spontaneous desires as more favourable to their freedom and self-rule than second-order ones. We feel that, in this case, it is second-order attitudes that should conform to first-order ones, and not conversely.

In spite of its possible shortcomings, Frankfurt's approach to the control or self-determination condition in terms of hierarchical and reflective attitudes has importantly shaped the discussion on free will and moral responsibility, as has also his influential criticism of the alternative possibilities condition. Compatibilist reflection on freedom

and moral responsibility is indebted, in several ways, to Frankfurt's criticism of earlier compatibilist thinking, as we shall try to show.

So one thing the preceding critical considerations may suggest is that second-order volitions are not the right or the only kind of reflective attitudes that are relevant to moral responsibility. Values should also enter into the picture. Watson has chosen this line. But those considerations may also suggest, in addition, that a satisfactory conception of control and moral responsibility should not be purely formal or structural, but should also take into account the content of the relevant attitudes and values. However, there is nothing to prevent compatibilists from including this content in their accounts of moral responsibility, and in fact some compatibilist approaches, such as Susan Wolf's, have insisted on this aspect.

For what concerns Brave New World cases, compatibilists are not defenceless against them. One possibility for them is to bite the bullet: to insist that causal origin is not relevant to self-determination and moral responsibility and to accept that Brave New World citizens are self-determined and morally responsible agents, explaining away our contrary intuitions. This seems in fact to be Frankfurt's choice when he holds that, even if a Devil or a neurologist provides a subject "with a stable character or program" which determines "the subsequent mental and physical responses of the subject to his external and internal environments", there are no reasons to deny that this subject may act freely or be morally responsible for what he does (cf. Frankfurt 1975:53). But compatibilists can also, and more plausibly, hold that our intuitions about Brave New World cases are correct and go on to argue that what explains these intuitions is not that Brave New World citizens lack libertarian freedom-related properties, such as ultimate control over their decisions and actions, but that they lack properties that can in principle be accounted for in compatibilist terms, such as appropriate sensitivity and responsiveness to reasons, including moral reasons, or an unimpaired capacity for practical reasoning. In addition to this, compatibilists may also put forward historical conceptions of moral responsibility and account for the relevant causal origin of our actions in a way consistent with determinism. Fischer and Ravizza are plausibly interpreted as having taken this dual line.

Value-based accounts of control: Watson, Wolf

Gary Watson shares with Frankfurt a dissatisfaction with the crudity and simplicity of classical compatibilism's psychological view of the self and its account of freedom and control. Classical compatibilism analyses freedom on the exclusive basis of desires, as an ability to do what one wants to do. But there are cases of acting on one's desires, and getting what one wants, which we would not count as cases of free action, though the agent's behaviour is intentional. This probably happens in addictive or phobic behaviour. So Watson rightly points out that classical compatibilist accounts of freedom "would seem to embody a conflation of free action and intentional action" (Watson 1982:97). However, there is nothing in compatibilism that forces it to work with such crude pictures of agents and their freedom. Watson, like Frankfurt, intends to remain faithful to compatibilism while remedying the defects of its classical versions.

While Frankfurt's view of control rests on a hierarchy of desires, so that an agent can be said to control her action, and the desire on which she acts, in so far as this desire (her will) is backed by a second-order volition, Watson construes the notion of control on the basis of a distinction between different and independent sources of motivation, namely desires and values. It is one thing to want or desire something; it is a different thing to value it or to judge it good or worthwhile. However,

> …to think a thing good is at the same time to desire it (or its promotion). Reason is thus an original spring of action. It is because valuing is essentially related to thinking or *judging* good that it is appropriate to speak of wants that are (or perhaps arise from) evaluations as belonging to, or originating in, the rational (that is, *judging*) part of the soul; values provide *reasons* for action. The contrast is with desires, whose objects may not be thought good and which are thus, in a natural sense, blind or irrational. Desires are mute on the question of what is good.
>
> (Watson 1982:99)

There are, then, "rational" wants, which are or arise from evaluations, and "mere" desires, which are independent of those evaluations. Watson thus rejects a Humean view of practical reasoning and motivation, according to which Reason is motivationally inert, and only passions (desires, in Watson's terms) can move an agent to act, in favour of what he calls a "Platonic" perspective, which holds Reason to be a motivating force itself. It is worth pointing out, however, that a rejection of Humeanism would need some more work, for it is open to the Humean to agree that rational evaluations can motivate while insisting that, if they do, it is only because they connect with pre-existing and Reason-independent desires. To reject this version of Humeanism, as represented, for example, by Bernard Williams's theory of internal reasons, is a notoriously difficult task, which Watson certainly does not undertake in his paper. Let us accept, however, for the sake of argument, that Watson is right in taking values to be motives independent of desire. On this assumption, there is a potential conflict between values and desires, and this conflict underlies and helps us to understand the problem of freedom: "The problem of free action arises because what one desires may not be what one values, and what one most values may not be what one is finally moved to get" (Watson 1982:100).

Whereas classical compatibilism conceives the self as centrally constituted by an agent's (non-pathological) first-order desires, and Frankfurt views it as essentially formed by an agent's second-order volitions, in Watson's perspective it is values that are to be identified with the self. This is why agents can be estranged from some of their desires and "be motivated to act in spite of themselves" (Watson 1982:102). Corresponding to these different views of the self, we can find different conceptions of what self-determination and control amount to. Whereas, in classical compatibilism, to act in a controlled and self-determined fashion is to do what one wants to do, because one wants to do it, and in Frankfurt's view it is to act on a desire that one wants to act on, because one wants to act on it, in Watson's account it is to be moved to act by one's values or, as he also puts it, by one's valuation system: "The free agent has the capacity to translate his values into action; his actions flow from his evaluational system" (Watson 1982:106). In Frankfurt's view, control centrally involves a harmony between effective first-order desires (the will) and second-order volitions. In Watson's perspective, it involves a

harmony between the agent's valuational and motivational systems, between what she values and what actually moves her to act. In both cases, the agent enjoys self-determination, in that it is her own self (her second-order volitions or her own values) that determines her actions.

Unlike Frankfurtian second-order volitions, Watson's evaluations are primarily first-order, that is, they are mostly about courses of actions: "Initially, [agents] do not (or need not usually) ask themselves which of their desires they want to be effective in action; they ask themselves which course of action is most worth pursuing" (Watson 1982:109). There is a connection between valuations and second-order volitions, "for the same considerations that constitute one's on-balance reasons for doing some action, *a,* are reasons for wanting the desire to do *a* to be effective in action, and for wanting contrary desires to be ineffective. But in general evaluations are prior and of the first order" (Watson 1982:109). Second-order volitions result from evaluations and are, in this sense, secondary with regard to them.

In spite of differences in content, Watson's theory of control is structurally very similar to Frankfurt's and, in our view, faces very similar problems as well. Let us look at some of them.

As in Frankfurt's case, it can be argued that control, as Watson conceives of it, is not necessary for moral responsibility. Though, in the paper we are analysing, Watson addresses the problem of freedom, and not of moral responsibility, his views undoubtedly have a bearing on the latter. Dealing with the question of compulsive agents, such as kleptomaniacs, Watson, like most of us, holds their thieving actions to be unfree. Presumably, he would also deny that they are morally responsible for them. His explanation of these judgements rests upon his view of freedom and control:

> What is distinctive about such compulsive behaviour, I would argue, is that the desires and emotions in question are more or less radically independent of the evaluational systems of these agents. The compulsive character of a kleptomaniac's thievery has nothing at all to do with determinism… Rather, it is because his desires express themselves independently of his evaluational judgements that we tend to think of his actions as unfree.
>
> (Watson 1982:110)

However, to see that this explanation cannot be completely right, we can resort to the example of Harry, the envious student, which we used to question the necessity of Frankfurtian control for moral responsibility. Remember that Harry wants Peter to fail in the next exam and, fabricating an excuse, refuses to lend him an important book. We may assume that Harry's values do not harmonize with his envious desires, which express themselves "independently of his evaluational judgements". We may even assume that Harry's valuational system does not approve of his refusal to lend the book to Peter. None the less, we judge that he acted freely and with the relevant control over his action, and that he is morally responsible for it. What explains our judgement that the kleptomaniac is not free and morally responsible for his thievery cannot be the discrepancy between his desires and his evaluations, for, if it were, we should also judge that Harry is not free and morally responsible for not lending the book to Peter.

This suggests that control, as Watson conceives of it, is too strong a requirement for freedom and moral responsibility, and not really necessary for them. Paraphrasing Frankfurt, we can express Watson's conception of control as follows: an agent exercises control over her actions by securing the conformity between her motivational and her valuational systems. Now, though Harry controls his action and is morally responsible for it, he exercises this control in the opposite way, namely by keeping his motivational and his evaluational systems apart, and indulging in acting on desires that he does not value. We judge that he acted freely and is morally responsible for what he did because we think that he controlled his action, in the sense that it was within his power to act on his values and lend the book to Peter. The concept of control we are employing here is what Kane calls "dual control" and Fischer and Ravizza "regulative control"; it is the sort of control that involves alternative possibilities. This is precisely the sort of control that we think the kleptomaniac lacks, and this, it seems, is why we do not hold him free and morally responsible for thieving. Unlike Harry, the kleptomaniac does not control the discrepancy between his values and his desires: he could not have acted on his values and avoided thieving. If this explanation is correct, the control over one's actions that is relevant to moral responsibility includes the availability of alternatives. *Pace* Watson, the compulsive character of the kleptomaniac's thievery may have to do with the question of determinism.

The discussion of these cases also suggests that the self is not to be identified with an agent's values (or, *mutatis mutandis,* with second-order volitions), for Harry can, willingly and consciously, distance himself from his values and keep them apart from his desires, allowing the latter to be effective in his action in spite, not of himself, but of his values or valuational system. In not lending the book to Peter he acts in a self-determined way, but his action is not determined by his values. He puts them, so to speak, in brackets, and goes ahead with his desires. The self, then, seems to be something behind both desires and values.

This is not meant to deny that Watson, and Frankfurt, have indicated aspects that are relevant and important for an understanding of free will and moral responsibility. Having a valuational system is plausibly held to be a requirement for both free will and moral responsibility. It may also be a requirement to have reflective, second-order volitions, though Watson seems to me right in holding the latter to be dependent on valuations and, in this sense, secondary with respect to them. Against classical compatibilism, Watson, and Frankfurt, have rightly insisted on the psychological complexity that agents should have in order to qualify as free and morally responsible. Reflexivity seems clearly to be required for this qualification. Both second-order volitions and values are or can be used as reflective attitudes towards primary motives or desires, and not only towards courses of action. However, the way these authors conceive of the control condition for moral responsibility, namely as an effective matching between motives and reflective attitudes, clearly seems too strong. As Harry's case shows, an agent can act with the relevant degree of control even if her desires and her values (or her second-order volitions) come apart. What is needed, it seems, is something weaker than effective matching, namely the capacity or ability to achieve that matching. This ability is, apparently, what Harry has and the kleptomaniac lacks. But, in this modalized formulation, control would seem to include alternative possibilities.

Watson's value-based account of control also faces the problem of Brave New World cases. Brave New World citizens can satisfy Watson's view of control, even if their valuational system is the product of a process of genetic manipulation, conditioning or brainwashing. If they cannot be said to be free and morally responsible agents, this undermines the sufficiency of Watson's construal of the control condition as well. In general, our remarks about Frankfurt's position in connection with this problem can also be applied to Watson's. And again, what seems to be at stake here is the question of ultimacy. Watsonian control is not deep enough to ground moral responsibility, and what it seems to be absent in it is the agents' control over the springs of their actions, including their values, that is, ultimate control.

As a purely structural or formal approach, with no restriction over the substance of an agent's valuations, Watson's theory also parallels Frankfurt's in facing a problem about content. Watsonian control requires desires to conform to values, for a free agent's actions "flow from his evaluational system" (Watson 1982:106). According to Watson, there is no reason why, among our other desires, second-order volitions should be privileged in being especially "our own". Watson's resort to valuations is supposed to meet this objection. But the objection can also be raised against Watsonian valuations as well. It is not difficult to think of situations, such as certain oppressive religious or social settings, in which an agent's spontaneous desires clearly seem to be more her own and favour her autonomy more than her values. In these contexts, actions that flow from the agent's valuational system can rightly be held to be less free than those arising out of her natural, spontaneous desires. In an extreme case, an agent can value her own enslavement and lack of freedom. It would be perverse to say that actions flowing from this evaluation are expressions of her free agency and self-determination.

Though the problem about content and the problem about ultimate control are different, there are, however, no sharp limits between them, for the former can also be related to the question whether, and to what extent, an agent's motives and values can be said to be really her own and have their ultimate source in herself.

Susan Wolf's work *Freedom Within Reason* (Wolf 1990) is an explicit attempt to lay a foundation of freedom and moral responsibility without resorting to the requirement of ultimate control, a requirement (which she calls "autonomy") that she deems incoherent: "The concept of an autonomous agent may seem to be an impossible one" (Wolf 1990:13). Wolf classifies the compatibilist accounts of control we have been looking at (classical compatibilism, Frankfurt, Watson) as versions of a non-autonomous conception of free will and moral responsibility, which she labels "The Real Self View". According to this conception, free agency and control, in the sense relevant to moral responsibility, amounts to one's actions being determined by one's (real) self, which, depending on the different versions, is centrally identified with one's (first-order) desires, second-order volitions or valuational system. The Real Self View is a non-autonomous conception in that it does not require that the self be an ultimate or causally undetermined source of one's actions or that it have ultimate control over them. For this view, questions about the self's origin are irrelevant to the issue of moral responsibility; this is, in our own terms, an ahistorical conception of control and self-determination: "This view does not require an agent to be endlessly accountable to herself… It is required that an agent *have* a real self, and that she be able to govern her behavior in accordance with it. But it does not matter where the real self comes from, whether it comes from somewhere else or

from nowhere at all" (Wolf 1990:35). It is scarcely plausible to think that defenders of the Real Self View will seriously consider the possibility that the self comes from nowhere. It would be more accurate to say that, for them, the possibility that the self is causally determined does not affect the agent's control over her actions or her moral responsibility for them. In this sense, its having no cause (implausible as this may be) would not affect these properties.

The Real Self View is, however, less than satisfactory as an account of freedom and moral responsibility. Wolf's objections against this view are quite close to those that a defender of ultimate control would tend to raise. According to Wolf, even if a subject's actions come from her real self, we can in some cases question her responsibility for those actions, "for we sometimes have reason to question an agent's responsibility *for* her real self. That is, we may think it is not the agent's fault that she is the person she is" (Wolf 1990:37). These cases may include some possible forms of mental illness in which "the self with which the victim completely and reflectively identifies is a self that other persons reasonably regard as being drastically mentally ill" (Wolf 1990:37), as well as some forms of psychological conditioning that have permanent effects on an agent's self. More common, and perhaps also more disturbing, are some cases of persons whose self and values are the product of a deprived or severely traumatic childhood, "persons who have fully developed intelligences and a complete, complex range of psychological structures, levels, and capacities for judgement, but who nonetheless do not seem responsible for what they are or what they do" (Wolf 1990:37).

These Wolfian objections seem to point to a demand for ultimate control and responsibility. They clearly show that the Real Self View does not offer a satisfactory and deep enough account of control and moral responsibility. They are closely related to some of the objections that we raised against Frankfurt's or Watson's proposals. And they would surely be endorsed by incompatibilist defenders of ultimacy requirements and seen by them as confirming the necessity of such requirements for moral responsibility. However, in a brilliant and provoking turn, Wolf rejects this necessity and holds that an alternative, non-autonomous conception, which she calls the "Reason View", can deal with these objections and provide necessary and sufficient conditions for moral responsibility. Lack of ultimate control is not the only explanation of our judgements about the problematic cases involved in the above objections. The Reason View can also account for them, with the crucial advantage of not resorting to demands for ultimacy of source and control, which, as we have suggested, Wolf takes to be impossible to satisfy.

The Reason View includes, among the conditions of free will and moral responsibility, a distinctive normative demand that adds to the requirements of Real Self views and distinguishes it from them. Wolf characterizes her view, by comparison with the Real Self approach, as follows:

> According to the Real Self View, an individual is responsible if and only if she is able to form her actions on the basis of her values. The Reason View insists that responsibility requires something more. According to the Reason View, an individual is responsible if and only if she is able to form her actions on the basis of her values *and* she is able to form her values on the basis of what is True and Good.
>
> (Wolf 1990:75)

Though the way Wolf pictures the Real Self View fits Watson's approach more exactly than those of classical compatibilism or Frankfurt, the term "values" can easily be interpreted, with no major distortion, to cover the latter positions as well. So the values on which a subject acts have to be objectively valid or correct for her to be rightly held morally responsible. There is, then, in Wolf's view, a double level of control: actions must be controlled by the agent's values (her self) and values (the self) must be controlled by what is objectively True and Good. This second level of control is control of the self by something other than itself, namely objective values, and is thereby a non-autonomous requirement. This feature clearly distinguishes Wolf's account from Autonomy views. Moreover, Wolf insists on the normative character of the extra condition she adds to those present in Real Self views, in contrast with the metaphysical nature of the requirement of ultimate control in Autonomy views. Finally, the Reason View shares with the Real Self View a stress on the irrelevance of the causal origin of an agent's values for what concerns her freedom and moral responsibility. Paraphrasing Wolf herself, what matters is that the agent has values and that these have been formed on the basis of the True and the Good, thus being objectively correct; it does not matter where these values come from.

We can now come back to the cases that seem intractable to the Real Self View (or, as we might also call it, pre-Wolfian compatibilism), such as those involving victims of psychological conditioning or of a deprived or severely traumatic childhood. Agents in these situations are rightly seen not to be (at least fully) morally responsible for their deeds, contrary to the judgement about them that follows from Real Self views. It can now be seen how the Reason View can provide an explanation of those cases that does not resort to the necessity of ultimate control and is, therefore, alternative to that provided by Autonomy views. According to the Reason View, the problem with agents in those problematic situations has nothing to do with the origin of their values, with those values having a source that is heteronomous or otherwise alien to the agents' selves. It has rather to do with the fact that, owing to the nature of their circumstances, they lack the capacity to form objectively correct values, or, to use Wolf's own terms, to form their values on the basis of the True and the Good. So, concerning victims of deprived childhood, Wolf writes: "A victim of a deprived (or depraved) childhood, for example, may be as smart as a person raised in a more normal environment, but, because of a regrettably skewed set of experiences, her values may be distorted. She is able to reason, as it were, but not able to act in accordance with Reason" (Wolf 1990:75–6). A similar incapacity would be involved in cases of psychological conditioning, mental illness and the like, which explains our judgements that those agents are not (fully) morally responsible. Moreover, although, in connection with psychological conditioning, Wolf refers to George Orwell's *1984* rather than to Aldous Huxley's novel, those situations which, following Watson, we labelled "Brave New World cases" would clearly belong to this group.

If the preceding explanation is correct, these problematic cases do not *ipso facto* vindicate the necessity of ultimate control and self-determination, and this is good news for moral responsibility, since, according to Wolf (and other thinkers), those requirements cannot possibly be met. Moreover, unlike those requirements, it would seem that there is nothing in a deterministic world that would necessarily preclude satisfying Wolf's normative condition. Thus one can rightly consider the Reason View as a form of compatibilism, as Wolf does herself.

It is, however, far from clear that Wolf's explanation of these cases is completely correct and that it can justify rejection of premise 1 of the incompatibilist argument at the beginning of this chapter, that is, of the natural assumption that ultimate control is a necessary condition of moral responsibility, understood as true desert. In our critical assessment of Frankfurt's and Watson's conceptions, we have distinguished two problems which, related as they may be, are none the less distinct, namely the problem of content and the problem of Brave New World cases, as we labelled them. The first problem has to do with the fact that agents can sometimes hold values whose content stands in the way of their freedom and (at least full) moral responsibility. An extreme example, we may recall, is an agent's valuing enslavement, including her own. More common cases include people raised in oppressive social or religious settings, holding values that run wildly against their own personal development as free and morally responsible agents. Now, it seems to me that Wolf addresses this problem plausibly. These agents do not satisfy the normative condition favoured by the Reason View: their values are not objectively correct; they have not been formed on the basis of the True and the Good. Though Wolf's theory of control and moral responsibility, like Frankfurt's or Watson's, is, as we have seen, ahistorical, in that, according to it, questions about (causal) origin are not relevant to moral responsibility, it is not, unlike these two, purely formal or structural. By virtue of its normative condition, Wolf's theory places restrictions on the content of agents' values. Not any value will do as a ground for moral responsibility.

It is less obvious, however, that Wolf's theory can deal successfully with the second problem, namely the problem of Brave New World cases. This problem would appear to do with the question of the source or origin of an agent's values rather than with the question of their content, and so connect directly with the question of ultimate control or, in Wolf's terms, autonomy. Wolf's strategy seems to be to reduce questions about the source or origin of an agent's values (and actions) to questions about the content of those values, or perhaps to eliminate the former in favour of the latter. Questions of origin are not relevant to moral responsibility. Provided that an agent can act on the basis of her values and form objectively correct values, she enjoys freedom and moral responsibility, no matter what source that ability and those values may have. The converse also holds. And, if the agent exercises that ability in acting, she acts freely and is morally responsible for what she does. And conversely.

Wolf's proposal may successfully deal with those cases where the crucial problem seems to be the content of the agent's values. It may also account for cases of deprived childhood: in many of these cases, the agent shows a deeply distorted, truncated or incoherent axiological system. But cases of psychological conditioning (Brave New World cases, more generally), though Wolf groups them together with those of deprived childhood, are resistant to this treatment. The apparent plausibility of Wolf's proposed explanation has to do with the fact that certain sorts of origins (deprived childhood, mental illness, psychological conditioning) of an agent's values usually result in a wrong or distorted system of values. But this is not necessarily so. It is conceivable that an agent forms her values through psychological conditioning and that these values are correct and not distorted at all. To see this, let us imagine that Brave New World is ruled by a team of benevolent, Platonic philosopher-kings, who know the True and the Good and have found the way to condition the citizens to form only objectively true values and to act on them. These agents would seem to match Wolf's necessary and sufficient conditions of moral

responsibility perfectly, and therefore they should be deemed fully morally responsible agents. But this consequence of Wolf's account is hard to accept. And what stands in the way of our accepting it seems to be closely related to the requirement of autonomy and ultimate control. There is, in this case, no problem about the content of values and their normative adequacy, or about the agents' ability to control their actions on the basis of those values. The only problem here seems to be the source of those values: they do not seem to be the agents' own, in the deep sense in which they should be in order for these agents to be justifiably held responsible and praiseworthy for what they do. These agents lack control over the values on which they act and this seems to be what accounts for our reluctance to accept that they are (fully) morally responsible and praiseworthy for their actions.

Though Wolf does not consider this precise case, it is quite plausible to think that her reaction to it would be to bite the bullet and insist on the moral responsibility of our Platonic New World citizens. In fact, at a certain point she writes: "If one is psychologically determined to do the right thing for the right reasons, this is compatible with having the requisite ability" (Wolf 1990:79). The ability in question is that which is necessary and sufficient for responsibility. However, according to Wolf, it is not compatible with having this ability, and so with moral responsibility, to be psychologically determined to act wrongly or for the wrong reasons, for in this case an agent cannot form her values on the basis of the True and the Good. But if these agents are not morally responsible, it is not because of psychological conditioning, but because of the wrongness of their values. Wolf's notorious asymmetry thesis about alternative possibilities follows from this contention: alternatives are required for an agent's moral responsibility for her morally bad actions, but not for her responsibility for her morally good ones. The upshot would seem to be, then, that, from Wolf's perspective, Platonic New World citizens would be morally responsible. However, if a New World is ruled by a team of wicked subjects, so that citizens are conditioned to have only morally wrong values and to act on them, then the citizens will not be morally responsible for their actions. These asymmetrical consequences of Wolf's approach strike me as wildly implausible. I would certainly agree that Wicked New World citizens are not morally responsible, but I would tend to express the same judgement about the Platonic New World inhabitants. There seems to be no basis for thinking that the former do not truly deserve blame for what they do while the latter do truly deserve praise. For what concerns their moral responsibility, their truly deserving blame or praise, there is no relevant difference between them, and the reason has to do with the ultimacy requirement: neither one nor the other has ultimate control over the springs of their actions, nor are they ultimate sources of either the springs or the actions. If someone deserves praise or blame for what these agents do, it is the respective teams of programmers.

That there is no difference in moral responsibility between Wicked New World and Platonic New World agents does not mean that there is no moral difference between them, for moral responsibility is not the only dimension on which people's worth and value can be assessed. Moral responsibility is not the only dimension of morality. And it is not self-evident that accepting the non-existence of moral responsibility would mean the end of moral values or of moral worth, though such acceptance would most probably lead to a deep revision of our view of the whole moral field, including our concepts of

moral values and worth. Now, even if neither Platonic New World nor Wicked New World citizens are morally responsible, in the sense of truly deserving praise or blame for their deeds and motives, the former might still be seen to be morally good and reliable people and to have a sort of moral worth that the latter would lack. In *this* sense, we might find the former, and not the latter, praiseworthy. But this sense of "praiseworthiness" is not the sense associated with the idea of moral responsibility for an action. It is more akin to the sense in which we praise someone for having certain qualities, such as elegance, insight or intelligence. The sense of praiseworthiness commonly associated with moral responsibility for an action involves the assumption that, the agent being the ultimate cause or origin of that action, which she could have avoided performing, she truly deserves praise for having performed it (and something similar would hold for blameworthiness). If this reflects our ordinary concept of moral responsibility, as I think it does, it is pretty clear that neither the Platonic nor the Wicked New World provides a foothold for it. And a conflation of those two senses of "praise" and "praiseworthiness" might help to explain the apparent plausibility of Wolf's presumable view that Platonic New World citizens would be as morally responsible and praiseworthy as ordinary agents are supposed to be.

Wolf might accuse us of committing a *petitio principii* against her, in assuming that ultimate control is a requirement for moral responsibility, which is the very question at issue. But we do not think that this charge would be justified. Our starting point has not been that the requirement is correct, but that it is natural for us to accept it, as it seems to be part of our ordinary concept of moral responsibility. And Wolf herself has readily acknowledged how natural that requirement is for us:

> It seems…that in addition to the requirement that the agent have control over her behavior (that she have a potentially effective will) and the requirement that she has control along the right lines (a relevantly intelligent will), there is a requirement that the agent's control be ultimate… This condition, like the others, seems to cohere with the meaning responsibility has for us… It makes sense that such beings should have a special significance, for they are sources of value (and disvalue) rather than mere carriers of it.
>
> (Wolf 1990:10)

This naturalness gives the requirement an initial presumption of truth. But we have not simply stopped here and transferred the *onus probandi* to those who reject it. We have positively argued for its truth, trying to show that ultimate control is, in effect, as important for moral responsibility as it initially seems to be, and that this requirement cannot be abandoned in favour of Wolf's normative condition. We have tried to show this by imagining a case (the Platonic New World) in which agents who plainly satisfy all the conditions proposed by the versions of compatibilism we have reviewed so far, including Wolf's, are not, however, morally responsible, and where the only thing that stands in the way of their moral responsibility seems to be their lack of ultimate control. This is not a definitive proof, for we cannot exclude the possibility that someone could come up with an alternative, and successful, explanation. But, in the absence of such an explanation, if there is a conflict between judgements on whether Platonic New World citizens are truly morally responsible, it seems to me that our strong pre-theoretical intuition that those agents lack moral responsibility, in the sense of true praiseworthiness, for what they

decide and do should be given pre-eminence over judgements induced by theoretical constructions that reject a natural assumption about moral responsibility.

Finally, Wicked and Platonic New World cases not only speak up for the requirement of ultimate control, but also for the view that ultimate control, as Kane holds, involves dual control, that is, available alternative possibilities and control over which of them is going to be actualized. Having already argued for the necessity of alternative possibilities for moral responsibility, we have no difficulty in accepting that this view, according to which there is a dependence relation between the alternative possibilities and the control conditions, may well be true.

Reasons-responsiveness and ownership: Fischer and Ravizza

John Martin Fischer and Mark Ravizza have recently developed a sophisticated and interesting account of moral responsibility which suggests an explanation of some of the problematic cases we have been considering that does not appeal either to alternative possibilities or to ultimate control. If this explanation can successfully account for all such cases, ultimate control could be shown not to be necessary for moral responsibility after all.

To be morally responsible, as Fischer and Ravizza propose to understand this concept, is to be an appropriate candidate for what, in an influential article, Peter Strawson called "reactive attitudes" (Strawson 1962:1–25), attitudes such as resentment, gratitude or indignation. Though they do not specify what being an appropriate candidate for such attitudes amounts to, this concept seems clearly related to the idea of truly deserving praise or blame, as well as to the justification of moral responsibility attributions. In requiring appropriateness, they are rejecting a purely pragmatic approach to moral responsibility and favouring a more realist perspective.

Fischer and Ravizza use "semicompatibilism" as a label for their general position. According to this position, determinism is not compatible with freedom to do otherwise, that is, with alternative possibilities being actually available to agents. But this does not imply that determinism is not compatible with moral responsibility, for the availability of alternative possibilities is not required for moral responsibility. In fact, they hold that determinism is compatible with moral responsibility, for the conditions that are necessary and sufficient for moral responsibility are compatible with determinism.

In a broad outline of this account, moral responsibility requires two general sets of conditions, which are also jointly sufficient for it, namely epistemic and freedom-relevant, or control, conditions. Epistemic conditions are succinctly characterized as follows: "An agent is responsible only if he both knows the particular facts surrounding his action, and acts with the proper sort of beliefs and intentions" (Fischer and Ravizza 1998:13). However, their main and almost exclusive focus is control, or freedom-relevant, conditions.

Fischer and Ravizza's proposal about control conditions of moral responsibility is deeply indebted to Frankfurt's criticism of the Principle of Alternate Possibilities (PAP) and to a reflection on Frankfurt cases. They are convinced by these cases that "moral responsibility does not require the sort of control that involves the existence of genuinely open alternative possibilities" (Fischer and Ravizza 1998:31). However, although,

according to them, Frankfurt cases show that *this sort* of control is not required for moral responsibility, they do not show that control is not required at all. Frankfurt cases feature a responsibility-undermining factor, but this factor operates only in the alternative sequence, not in the actual sequence, where it remains causally inert. By virtue of this causal inertness, the agent retains control over her deliberation and action in the actual sequence, though, given that she could not have done otherwise, this control does not extend over alternative actions. If, in these circumstances, the agent is morally responsible for her action, this points to a distinction between two sorts of control, namely the sort of control that an agent in a Frankfurt case enjoys in the actual sequence and the sort of control that she would enjoy if, in the absence of the counterfactual intervener, she could have done otherwise. Fischer and Ravizza call the first sort of control "guidance control", and the second "regulative control". They characterize these two types of control as follows:

> Let us suppose that Sally is driving her car… Insofar as Sally actually guides the car in a certain way, we shall say that she has "guidance control". Further, insofar as Sally also has the power to guide the car in a different way, we shall say that she has "regulative control"…
>
> When Sally freely acts so as to cause the car to go to the right, she exhibits guidance control. Guidance control of an action involves an agent's freely performing that action… Regulative control involves a *dual* power: for example, the power freely to do some act A, and the power freely to do something else instead…
>
> (Fischer and Ravizza 1998:31–2)

We naturally tend to think that having guidance control over one's actions implies having regulative control over them as well. But Frankfurt cases show that this natural assumption fails. Agents in Frankfurt cases have guidance control over their actions but they lack regulative control over them. So it is possible to have the first kind of control but not the second. Now if, in Frankfurt cases, the agent is morally responsible for her action while lacking regulative control, this implies that regulative control is not required for moral responsibility. It goes without saying that we do not agree that, in Frankfurt cases, agents lack—in Fischer's terms—regulative control, as we have tried to argue in the preceding chapter. But we are now trying to understand Fischer and Ravizza's proposal. Their suggestion is that guidance control, but not regulative control, is necessary and, provided that the cognitive conditions are met, also sufficient for moral responsibility. But what does guidance control consist in?

The thought that regulative control is required for moral responsibility may arise out of reflection on such cases as brainwashing, drug addiction, hypnosis or brain manipulation. In these cases, agents are not reasonably held to be morally responsible for their actions. What explains this judgement is the actual operation of a factor that undermines the agents' control over their actions by impairing their "mechanism" (the term is Fischer's) of practical reasoning and decision-making. According to Fischer and Ravizza, it is natural to think that the sort of control that is thereby being undermined is regulative control (freedom or ability to do otherwise). The point can be put in terms of sensitivity to reasons, including reasons to do otherwise. Agents in these cases lack appropriate sensitivity or responsiveness to reasons. A hypnotized agent, for example,

unlike an agent with unimpaired deliberative powers, "would still behave in the same way, no matter what the relevant reasons are… Thus it is very natural and reasonable to think that the difference between agents who are morally responsible and those who are not consists in the 'reasons-responsiveness' of the agents (and thus their possession of regulative control)" (Fischer and Ravizza 1998:36–7).

It is again reflection on Frankfurt cases that corrects this natural explanation of the cases at hand, for agents in Frankfurt cases lack reasons-responsiveness and regulative control in this sense and, none the less, are morally responsible. In order to account for this difference, it is important to focus on what happens in the actual sequence and to distinguish between agents and their mechanisms of deliberation and decision. So in cases of hypnosis or brainwashing the actual mechanism is not reasons-responsive, whereas in Frankfurt cases "the kind of mechanism that *actually* operates *is* reasons-responsive, even though the kind of mechanism that *would* operate—that is, that does operate in the alternative scenario—is *not* reasons-responsive… Thus, the actual-sequence *mechanism* can be reasons-responsive, even though the *agent* is not reasons-responsive. (*He* couldn't have done otherwise)" (Fischer and Ravizza 1998:38–9). That is, if we kept constant the actual deliberative mechanism of the agent in a Frankfurt case, she would appropriately respond to reasons for doing otherwise if such there were. It is, then, the difference between the reasons-responsiveness of the mechanisms that operate in the actual sequence, and not a difference in regulative control or ability to do otherwise, that accounts for the fact that hypnotized agents, unlike agents in Frankfurt cases, are not morally responsible.

So the sort of control that agents exhibit in Frankfurt cases involves reasons-responsiveness of the mechanism of deliberation and decision-making that operates in the actual sequence. This is, then, a first component of guidance control. But guidance control includes a second component, which Fischer and Ravizza also derive from reflection on Frankfurt cases, namely ownership of that mechanism. The mechanism that is at work in the actual sequence of a Frankfurt case can be rightly said to be the agent's own. In contrast, the mechanism that would be at work if the counterfactual controller were to intervene is somehow external to the agent and cannot properly be considered as her own (cf. Fischer and Ravizza 1998:39). Ownership is thus taken to be a second aspect of guidance control, conceptually independent of the first, or reasons-responsiveness. The whole notion of guidance control is summarized in the following text: "An agent exhibits guidance control of an action insofar as the mechanism that actually issues in the action is his own, reasons-responsive mechanism" (Fischer and Ravizza 1998:39). Let us now look in more detail at each of the two components of guidance control, starting with reasons-responsiveness.

Reasons-responsiveness can be had in different degrees. An agent enjoys it in its highest degree (she has "strong reasons-responsiveness") just in case, if the "mechanism that actually issues in the action…were to operate and there were sufficient reason to do otherwise, the agent would *recognize* the sufficient reason to do otherwise and thus *choose* to do otherwise and *do* otherwise" (Fischer and Ravizza 1998:41). In these conditions, the agent is maximally sensitive or receptive to reasons and maximally reactive to them: in all possible worlds in which the actual mechanism operates and in which there is sufficient reason to do otherwise, the agent recognizes that reason and chooses to do and does otherwise. However, desirable as strong reasons-responsiveness

may be, cases of weak-willed agents, for example, show that it is too demanding a condition and not really necessary for moral responsibility.

A lower degree of reasons-responsiveness, which Fischer and Ravizza call "weak reasons-responsiveness", may be all that is required for moral responsibility: "Under *weak reasons-responsiveness,* we (again) hold fixed the actual kind of mechanism, and we then simply require that there exist *some* possible scenario (or possible world) in which there is a sufficient reason to do otherwise, the agent recognizes this reason, and the agent does otherwise" (Fischer and Ravizza 1998:44). However, to avoid some counterexamples to this formulation, they add some conditions to weak reasonsresponsiveness. This yields "moderate reasons-responsiveness", their final considered version of the reasons-responsiveness component of guidance control. Moderate reasons-responsiveness is just weak reasons-responsiveness with some additional features: an appropriate connection between reasons to do otherwise and doing otherwise, a variable response of the mechanism according to the strength of those reasons, an intelligible pattern in that response, and responsiveness not only to prudential reasons, but to some moral reasons as well.

It is quite reasonable to think that something like this condition is required for moral responsibility. If, with her actual deliberative "mechanism" in play, the agent persists in acting as she does no matter how strong the reasons for doing otherwise, this clearly suggests that her mental abilities are not working well enough to hold her morally responsible for what she is actually doing. Moderate reasons-responsiveness, then, involves the truth of some counterfactuals: it is not only a question about how the agent does actually choose and act, but also about how she would choose and act if certain circumstances were to hold. There is, however, no suggestion that the agent must have access to those possible worlds in which she chooses and acts otherwise, or that those alternative ways of acting have to be actually available to her: regulative control over alternative possibilities is not required. The account remains an "actual-sequence" approach to moral responsibility: moral responsibility depends on how things are in the actual sequence that leads to the action. The truth of those counterfactuals is required to establish that the deliberative and decision-making mechanism that operates in the actual sequence is sound enough and gives the agent a sufficient degree of control over what she actually chooses and does to ground her moral responsibility for it.

It is worth noting, however, that, though actually available alternative possibilities are not required for moderate reasons-responsiveness, the *notion* of alternative possibilities is essential to it. Fischer and Ravizza's reasons-responsiveness condition can be seen as a refined version of a conditional analysis of the alternative possibilities requirement for moral responsibility. In fact, it looks close to Davidson's version of this analysis, where the conditions under which the agent would do otherwise are reasons, and not decisions, tryings or acts of will in general, as happens in older versions. Conditional analyses are attempts to account for the intuitive idea that moral responsibility requires the ability to do otherwise in a way that is consistent with determinism, and this is also the case with moderate reasons-responsiveness. Determinism may exclude that alternative possibilities be effectively available to agents, but this availability (the agent's volitional control over her access to those possible worlds in which she chooses and does otherwise) is not required for moderate reasons-responsiveness.

Let us now look at the second component of guidance control, namely ownership of the mechanism that actually issues in the action. If reasons-responsiveness can be said to be Fischer and Ravizza's compatibilist counterpart to the alternative possibilities requirement for moral responsibility, ownership would seem to be their answer, again in compatibilist spirit, to the source problem, with its associated requirement of ultimate control. Unlike Frankfurt, Watson or Wolf, Fischer and Ravizza address this problem in a direct way, instead of eliminating it or reducing it to structural or content problems. Remember that, for Frankfurt, the origin of an agent's second-order volitions and their matching with first-order desires, as well as of her identifications with either, was not relevant to her moral responsibility. Watson and Wolf adopt a similar position in their respective proposals about the conditions of moral responsibility. This point can be expressed by saying, as we did, that these theories are ahistorical. In contrast, Fischer and Ravizza's approach to moral responsibility is explicitly historical. It is the second component of guidance control that gives this approach its historical character, for the properties that constitute reasons-responsiveness are, so to speak, "snapshot" properties, in that having them does not depend on their origin: "The properties in virtue of which a mechanism…is reasons-responsive are modal or dispositional properties of an agent *at (roughly) the time of the action*… But…we shall argue for the claim that moral responsibility is an essentially historical notion" (Fischer and Ravizza 1998:170).

The historical component of guidance control is conceptually independent of reasons-responsiveness: the question whether, and to what degree, an agent enjoys reasons-responsiveness does not imply anything about the origin of her deliberating and decision-making mechanism, and vice versa. Since both components are required for guidance control, an agent can have full reasons-responsiveness but lack guidance control, and therefore moral responsibility, owing to the history of her practical reasoning mechanism. Fischer and Ravizza, then, are aware of the problem that source questions raise for compatibilist accounts of moral responsibility, a problem that, under the guise of the ultimate control condition, we have seen to arise for all compatibilist approaches discussed in this chapter. These approaches ground moral responsibility in the agent's possession of a certain psychological structure, with certain specified relationships among its elements. In Frankfurt's theory, for example, the psychological structure has to include second-order volitions, as well as conformity (or "mesh", as Fischer and Ravizza put it) between them and the agent's will, or effective first-order desires. Questions about the origin of these features are deemed utterly irrelevant to moral responsibility. Fischer and Ravizza include Hume's and Watson's in the group of "mesh" theories. Wolf's view could plausibly be added to them. All these theories, as we have seen, stumble on the problem of the origin or source of the favoured psychic structure. As Fischer and Ravizza point out,

> …these features—the selected "mesh" or configuration of internal mental states—can be induced by such processes as hypnosis, brainwashing, and even direct stimulation of the brain… These cases rather graphically illustrate the intuitive idea that what matters, for moral responsibility, is not just the arrangement of mental ingredients, but how that arrangement is produced.
>
> (Fischer and Ravizza 1998:187)

It is quite clear that Brave New World cases also belong to this group of responsibility-undermining historical or genetic processes. As we have seen, Fischer and Ravizza do not think that the historical, or source, problem can be solved in purely structural and non-historical terms, as Frankfurt, Watson or Wolf try to do. It has to be addressed as such, by specifying the historical conditions that have to be met in order for an agent's deliberative mechanism, and the ensuing actions, to be truly *her own,* so that one can justifiably judge *her* to be morally responsible for those ensuing actions.

Fischer and Ravizza's specification of the historical conditions for the mechanism that issues in an agent's actions to be her own takes the form of a description of the process by which a child becomes a moral agent:

> It is perhaps useful to divide this process into three stages: "training", "taking responsibility", and "being held responsible". When parents treat their child as if he is responsible—taking certain attitudes toward him—they are engaged in moral *training*. This sort of training aims to induce a certain sort of view in the child, a view of himself as an agent and, in some situations, a fair target for praise and blame. Adopting this view is an important part of *taking responsibility*. Having adopted this sort of view, the child is ready to be *held responsible* (by others as well as himself)… The different stages overlap with one another and compose a cycle that must be continually repeated in the process of becoming a fully morally responsible agent.
>
> (Fischer and Ravizza 1998:210)

A crucial step in this process is that of taking responsibility. Fischer and Ravizza conceive of this step, rather than as an action, as the acquisition by the subject of a certain array of beliefs: that she is an agent, whose choices and actions have effects on the world; that if she chose and acted in a different way, the effects on the world would be different as well; that she is a fair target for the reactive attitudes. Moreover, these beliefs have to be appropriately based on the evidence. Through this process, the subject takes responsibility for the springs of her actions and makes them, and the ensuing actions, her own.

This process, together with moderate reasons-responsiveness, gives the complete content of the condition of guidance control, which is both necessary and (given the cognitive requirements) sufficient for moral responsibility. When both the historical condition and the requirement of moderate reasons-responsiveness are met, the agent has guidance control over her actions and is (given the cognitive requirements) morally responsible for them. Fischer and Ravizza extend this account of responsibility for actions, with some qualifications, to cover responsibility for omissions and consequences as well.

Fischer and Ravizza's theory of moral responsibility is probably the most systematic and refined attempt, so far, to account for moral responsibility in a compatibilist frame. This theory incorporates many virtues of previous compatibilist approaches, while avoiding some of their shortcomings. Like Frankfurt's, it is an actual sequence theory; it is not, however, unlike Frankfurt's, Watson's and Wolf's, purely structural, but historical; and, like Wolf's, but unlike Frankfurt's and Watson's, it is not purely formal, for it includes some content features in the conditions of moral responsibility, by requiring the agent's responsiveness to (at least some) moral reasons. Like all actual-

sequence theories, it does not require the availability of alternatives (regulative control) for moral responsibility, so that it can resist challenges based on the incompatibility between determinism and freedom to do otherwise. Its two essential requirements, namely that the mechanism that actually gives rise to action be moderately reasons-responsive and that it be the agent's own, provide it with important resources to explain problematic cases within the limits of a compatibilist perspective. Let us focus on some of its explanatory virtues.

This account is not worse than Frankfurt's, Watson's or Wolf's in its ability to explain why, in some cases of drug addiction, phobias or kleptomania, agents are not morally responsible. The reason, according to this theory, is that, in those cases, the agents' actual mechanisms of deliberation and decision-making are not moderately reasons-responsive. However, this theory is in a better position than Frankfurt's or Watson's to explain the difference between our judgement about these cases and cases in which the agent acts out of motives that she does not approve of, as happens in our example of Harry, the envious student, who refuses to lend a book to Peter, a brilliant colleague of his. The theory might explain this difference by pointing out that Harry's actual mechanism of practical reasoning, but not the drug addict's, is reasons-responsive: in some scenarios, keeping the actual mechanism constant, Harry would recognize sufficient reasons to lend the book to Peter and would do so. An important advantage of this explanation is that it is a plausible alternative to an explanation in terms of a difference in regulative control (alternative possibilities) between the two agents. So it remains firmly on a compatibilist ground.

Fischer and Ravizza consistently reject regulative control as a requirement for moral responsibility, whereas Wolf, as we know, adopts an asymmetrical stance regarding it: alternatives are required for moral responsibility in cases of morally wrong actions, but not in cases of morally good ones. According to Wolf, agents who are determined to act rightly and for the right reasons are morally responsible, even if they are not able to do wrong, since they plainly have the required ability, namely the ability to act on their values and to form these values on the basis of the True and the Good, but agents who are determined to act wrongly are not morally responsible, because they lack that ability. According to Fischer and Ravizza, it all depends, in both cases, on whether the agent enjoys guidance control, that is, whether the mechanism that operates in the actual sequence is both the agent's own and moderately responsive to reasons. Now, remember that moderate reasons-responsiveness includes some responsiveness to moral reasons. On this basis, the theory's verdict about agents determined to act wrongly would presumably be the same as Wolf's, namely that they are not morally responsible. Concerning the other case, Fischer and Ravizza might also agree with Wolf on the agents' moral responsibility, provided that the agents meet the requirements for guidance control. However, Fischer and Ravizza's explanation seems to be preferable to Wolf's in that it gives a unified account of responsibility for morally good and bad actions and does not resort to the alternative possibilities condition in either case. This feature gives the theory a more consistent compatibilist structure.

However, by virtue of its much looser restrictions about content, this theory fares worse than Wolf's in cases where the crucial question seems to be precisely the content of an agent's attitudes and values. Even in the extreme case of someone who values enslavement, including her own, Fischer and Ravizza's verdict, unlike Wolf's, might be that the agent is fully morally responsible, provided that she enjoys guidance control.

Something similar could be said about agents raised in some oppressive social, religious or political environments. Wolf's strong normative requirement for moral responsibility gives her theory an advantage over Fischer and Ravizza's in explaining our intuitions about cases of this sort.

Concerning cases of deprived childhood, both theories might yield the same negative assessment about the agent's moral responsibility, though for different reasons. Wolf would insist on the distortion of the agent's values. As for Fischer and Ravizza, since an agent in those circumstances can satisfy the requirement of moderate reasons-responsiveness, they would have to appeal to the ownership component of guidance control and say that, if the agent is not morally responsible, this is owing to the fact that, by virtue of the circumstances of her upbringing, she has not gone through the process by which children become moral agents. Depending on particular cases, this explanation might be found wanting.

However, the crucial test for this theory, as for the rest of those we have discussed in this chapter, is cases in which the obstacle for the agents' moral responsibility seems to lie in the origin of their values and motivational system. Following Watson, we have labelled cases of this sort "Brave New World cases". Fischer and Ravizza's theory might seem to be better equipped to deal with these cases than the rest of the compatibilist theories we have examined, for it directly and consciously addresses the origin or source problem. The challenge for compatibilist accounts of control is that nothing short of an agent's being an ultimate source of her action, by having ultimate control over it, seems to be sufficient for moral responsibility ascriptions to be truly deserved and justified. Compatibilists must reject this claim, and insist that ultimate control is not actually required for moral responsibility, since it is not consistent with determinism. But none of the attempts to do without such a requirement that we have so far considered has proved successful, and the possibility of Brave New World cases has been the main reason for that failure. Let us now see whether Fischer and Ravizza's theory can do better.

We have already anticipated that this theory's verdict about a Platonic New World would be the same as Wolf's, provided that these citizens have guidance control. If this is actually so, then our criticisms of Wolf's proposal would also apply to Fischer and Ravizza's. But let us assume, for the sake of argument, that Platonic citizens do not comply with some aspect of guidance control. Suppose, for example, that they are not sensitive enough to prudential reasons to be said to enjoy moderate reasons-responsiveness. We can imagine, then, that the New World is not ruled either by benevolent Platonic philosopher-kings or by irredeemably wicked scientists, but by subjects who are somehow in between these two extremes and who are, in fact, more similar to present human beings in moral terms. Let us call this world Middle New World. Suppose now that citizens are conditioned so as to enjoy moderate reasons-responsiveness, including the requirements about their patterns of reasons-recognition, variation of response, and responsiveness to (at least some) moral reasons. Suppose that the ruling team decides about the content of their motivational system, about which values they will embrace and act upon and which reasons they will find convincing. Under these circumstances, it seems that, independently of what they themselves may feel or think, they are not morally responsible agents. The obvious root of this judgement would seem to be our intuitions about ultimacy of source and control: they are not morally responsible because they do not have ultimate control over the springs of their

actions nor can they be said to be their ultimate source. Compatibilists (and semi-compatibilists) must find an alternative explanation of that judgement. But now it seems that Fischer and Ravizza can provide this explanation.

According to Fischer and Ravizza's theory, agents who enjoy moderate (or even strong) reasons-responsiveness might still not be morally responsible, provided that they do not satisfy the second, historical component of guidance control, namely that the actual mechanism that issues in their actions be their own. And this is the likely reaction of the theory towards this sort of case, namely that the agents' mechanisms of deliberation and decision-making in Middle New World are not their own, and that this is why they are not morally responsible. But we cannot see how this reply can do. According to this theory, an agent's actual mechanism is her own provided that she has gone through a process that includes training, taking responsibility, and being held responsible. Taking responsibility, as we saw, is a matter of her coming to have a certain set of beliefs, appropriately based on the evidence, about her being an agent and a fair target for reactive attitudes. But now consider that, as part of the conditioning, agents may be led to train their children as moral agents, to acquire the set of beliefs included in their taking responsibility, and to be ready to be held responsible. So citizens of Middle New World can actually meet the conditions that, according to Fischer and Ravizza's theory, are both necessary and sufficient for being morally responsible agents. Their mechanisms can be said to be their own, in terms of this theory. But these consequences of the theory just seem plainly wrong. Ownership of the mechanism is too weak a surrogate of the ultimate control condition, and cannot do the job that it was meant to do.

A possible reply, on behalf of the theory, might be that these citizens' relevant beliefs are not appropriately based on the evidence, but this clearly need not be so. Provided that they are kept unaware of the actual process by which their motivational system and mechanisms of practical reasoning have been produced, their beliefs about their agential condition and so on can be fully justified, even though they are false. Bear in mind that, as far as we know, we might be such citizens. In the light of this, one might be tempted to reply that these citizens do not have true beliefs about the origin of their motivational system and of their mechanisms of practical reasoning, so that their taking responsibility for them is flawed. So having true beliefs about this matter might be included in the ownership condition. But this is clearly too strong a condition for moral responsibility. Remember that we do not know much about this question, which is a subject of much controversy in psychology, sociology, neurobiology, cognitive science and other disciplines. If it were really required, nobody (except perhaps those who actually knew what that origin is) could be morally responsible.

We are aware that not everybody will share the judgement that Middle (and other Brave) New World citizens are not morally responsible. At least some compatibilists will reject it. We think that spontaneous intuition supports this judgement, but intuitions and theoretical commitments are not always easy to distinguish: sometimes the former can mesh with the latter and not be taken by both sides to be neutral enough to decide about an issue. And this charge can be directed by incompatibilists at compatibilists and vice versa. In the case at hand, for example, compatibilists may hold that the judgement that these citizens are not morally responsible is unconsciously fed by incompatibilist inclinations, and incompatibilists, in turn, may hold that the opposite judgement is fed by compatibilist allegiances. It is important, then, to try to base one's position on intuitions

that are strong and shared by both parties. In Brave New World cases, as we have presented them, an obstacle may be that we have spoken in too loose and vague a way about conditioning. A natural way of conceiving this process is in terms of programming. In a slightly different context, where he imagines a Devil/neurologist (D/n), Frankfurt aptly describes this possibility: "D/n provides his subject with a stable character or program, which he does not thereafter alter too frequently or at all" (Frankfurt 1975:53). On this interpretation, some compatibilists may not have the intuition that subjects are not morally responsible. Frankfurt is among them. According to him, if this is how conditioning takes place, there are no decisive reasons for rejecting the agent's moral responsibility. In these circumstances, an agent…

…may become morally responsible, assuming that he is suitably programmed, in the same way others do: by identifying himself with some of his own second-order desires… And the person thereby *takes* responsibility for the pertinent first- and second-order desires and for the actions to which these desires lead him.

(Frankfurt 1975:53)

However, even on this interpretation of the conditioning process, it is hard to accept that Brave New World citizens are morally responsible. The problem with Frankfurt's proposal is that the subject's process of becoming morally responsible, through her identifying with, and taking responsibility for, her desires and actions, may clearly be itself part of the program. And, though Fischer and Ravizza understand "taking responsibility" in a slightly different way, this does nothing to prevent their proposal from facing this problem as well. As we described Middle New World, the full process of becoming a morally responsible agent, as Fischer and Ravizza conceive of it, can be something that citizens are conditioned to go through. And if these subjects are not even in control of the process by which they (supposedly) become morally responsible agents and take responsibility for their actions, it is really hard to accept that their degree of control over their actions is sufficient for them to be truly morally responsible for them.

But if this is not convincing enough, there is nothing to prevent us from interpreting the process of conditioning in a more direct and less polite fashion. We can imagine, for example, that the ruling team controls the citizens' mental lives by monitoring their brain states and directly manipulating them beyond the subjects' awareness, though with full respect for the conditions of guidance control. Under these conditions, few compatibilists, if any, will still insist that the subjects can be morally responsible. Derk Pereboom (cf. Pereboom 2001:120–1) has in fact argued that, in terms of Fischer and Ravizza's proposal, a subject whose brain is directly manipulated could count as a morally responsible agent. Their reply, as Pereboom reports it, is that this subject would not be a coherent self. This reply is, in fact, much the same as that which Frankfurt gives to a similar objection. As he puts it, if the manipulation takes place "on a continuous basis…so that each of the subject's mental states is the outcome of specific intervention…the subject is not a person at all. His history is utterly episodic and without inherent connectedness" (Frankfurt 1975:53). But, as Pereboom rightly points out, the mental life of a subject whose brain is being directly manipulated need not lack full connectedness and inner coherence. It could well be qualitatively identical to that of a non-manipulated person and so no less coherent than it. In Middle New World, if the

manipulation is intelligent and subtle enough, the citizens may not feel any disruption or incoherence in the unfolding of their conscious mental processes.

It seems, then, that, in spite of its virtues, Fischer and Ravizza's "semi-compatibilism" stumbles on much the same obstacle as the other compatibilist attempts we have examined in this chapter. And it clearly seems that what should be added to their otherwise rather plausible list of necessary conditions for moral responsibility, in order to obtain a sufficient condition for it, is, at least, the requirement of ultimate control, or true self-determination. On the assumption that alternative possibilities are themselves required for ultimate control, this condition might also be sufficient. "Ultimate regulative control" could be a suitable label for this enlarged condition.

Final remarks and Conclusion

In this chapter, we have mainly argued for the necessity of ultimate control for moral responsibility. We have started by giving some positive reasons for this contention, based on both the (partly) causal and the strongly evaluative nature of moral responsibility ascriptions, as well as on their deep effects on the addressee's personal worth. It clearly seems that we assume this ultimacy of source and control in ascribing moral responsibility to someone for something she did or caused to happen through her action. And we have then proceeded to defend the truth of that assumption in negative terms, namely by showing how accounts of moral responsibility which are intended to do without it do not succeed in providing a sufficient condition of moral responsibility. All such attempts face problems raised by Brave New World cases. These are examples of what Kane has called Covert Non-constraining Control (CNC) situations (cf. Kane 1996:64–71), in which agents judge, decide and act as the controllers want them to do, but without being constrained by them to do so. These cases point with particular clarity to the necessity of ultimacy of source and control for moral responsibility. Though we have gone through particular compatibilist attempts to dispense with this condition, it is unlikely that other, new attempts can succeed, for the examination of these particular accounts clearly suggests a unified strategy against any approach of this general sort. The strategy can be outlined, roughly, as follows: take any particular compatibilist approach and put its favoured conditions of moral responsibility in the context of a Covert Nonconstraining Control situation; the result will be that those conditions (which, obviously, cannot include ultimate control) do not suffice for moral responsibility; and, given that these conditions may include whatever cognitive, affective or volitive features of an agent are supposed to be compatible with determinism, the only remaining explanation is that the features that are required to have a sufficient condition are not compatible with it; this leaves, as plausible candidates, ultimate control and the availability of alternative possibilities; but Brave New World (or, more generally, CNC) cases are specifically designed to test the necessity of ultimate control; so, what these cases show is that, at least, ultimate control is required for moral responsibility, whether or not alternative possibilities are required as well. We think that the correct position is to

hold that the alternative possibilities condition is itself a component of the sort of control required for moral responsibility, and we have suggested "ultimate regulative control" as a suitable label for this unified condition. However, the necessity of the alternative possibilities condition for moral responsibility can be argued for along independent lines, as we have tried to do in the preceding chapter.

We have not argued in detail, however, for incompatibility between determinism and ultimate control. But this contention does not seem controversial. Not even compatibilists question it. As we suggested at the beginning of this chapter, it seems fairly obvious, given the general concepts of determinism and of ultimate control, that their instantiations cannot coexist.

Now, from the incompatibility between ultimate control and determinism, together with the necessity of ultimate control for moral responsibility, it clearly follows that, if determinism is true, moral responsibility is not possible, which is precisely SMR's premise B.

Before concluding this chapter, however, it may be worth pointing to a line of resistance to the sort of argument we have developed. Daniel Dennett (1984) has especially emphasized this line, which could be put as follows. What Brave New World (or CNC) cases may show is, at most, that manipulation by external *agents* deprives a subject of the control over her behaviour that is required for moral responsibility. It does not show that causal determination by *natural factors and laws* excludes such control as well. According to Dennett, what is doing all the work in this sort of argument for the incompatibility between determinism and moral responsibility is that the determining factor is manipulation by agents. But natural forces, not being agents, do not *manipulate or control* us, even if they causally determine our actions. So arguments that rest on manipulation or control by agents are flawed in that they fail to notice this important distinction.

This criticism is strongly reminiscent of that which Hume directed at those who held that causal necessity rules out freedom, namely that they fail to distinguish between causation and constraint. But neither is plausible. It is true that there are differences between human agents and blind natural forces, and between our decisions and actions having their origin in the ones or the others. But these differences, contrary to Dennett's contention, are not relevant to the question at issue. Even compatibilists, such as Watson, agree on this point: "My freedom to dance is equally impaired whether my legs are paralysed by organic disease or shackled by human hands. What needs explanation is that Brave New World individuals are impaired in certain ways. It is a mistake to think that it matters whether this impairment has a natural or human origin" (Watson 1987:151–2). In fact, we could replace the ruling team in a Brave New World by an impersonal mechanism without harming the conclusion about the agents' lack of moral responsibility. In featuring external controllers, Brave New World cases depict with special clarity a situation in which agents do not have ultimate control over their actions and cannot be considered as their ultimate sources. But this is also the case if determinism holds, for there is no place for ultimate sources or ultimate control in a deterministic world. Including external controllers is merely a didactic resource, not an essential ingredient of the argument.

In the light of all this, we can conclude that the argument that we have examined in this chapter provides premise B of SMR—namely that, if determinism is true, moral responsibility is not possible—with a second, and very strong, foundation. Together with the support derived from the argument about alternative possibilities that was examined in the preceding chapter, the truth of that premise does certainly appear to have very secure credentials.

4
Indeterminism and moral responsibility
(SMR's premise C)

We have so far gone through two arguments, each of which, if sound, already establishes SMR's premise B, that is, the incompatibility between determinism and moral responsibility. The first we labelled "the Incompatibilist Argument". To recapitulate, it is as follows: 1) Moral responsibility requires alternative possibilities: an agent is morally responsible for an action of hers only if she could have done otherwise. 2) Determinism rules out alternative possibilities: if determinism is true, nobody could have done otherwise than she in fact did. 3) Therefore, if determinism is true, moral responsibility is not possible. The second runs roughly as follows: 1) If determinism is true, ultimate control is not possible. 2) Ultimate control is necessary for moral responsibility. 3) Therefore, if determinism is true, moral responsibility is not possible. Both arguments are valid. And, in the preceding chapters, we have gone through a discussion of their premises. Chapters 1 and 2 dealt with the premises of the first argument and Chapter 3 with those of the second. The result of the discussion has been that there are important reasons for thinking that all such premises are true. Therefore we now have a strong case for the thesis that determinism is incompatible with moral responsibility. If this thesis is true, the only hope for the possibility of moral responsibility is that it be compatible with indeterminism. Premise C of SMR denies this compatibility. In the present chapter we shall be dealing with premise C. We shall see that there are important arguments and considerations that support it. If this premise is also true, then, provided that determinism and indeterminism exhaust the logical possibilities, the conclusion is scepticism about the possibility of moral responsibility.

Alternative possibilities, ultimate control and indeterminism

Two crucial steps in the arguments for incompatibilism were the premises stating the necessity of alternative possibilities and of ultimate control for moral responsibility. These are characteristic incompatibilist claims. Libertarian incompatibilists hold that these two conditions can (sometimes) be satisfied provided that indeterminism holds. There are certainly reasons for this contention. Concerning alternative possibilities, it clearly seems correct to assume that indeterminism (in the right places) allows for their availability to agents. But any reasonable incompatibilist must hold that alternative possibilities are only necessary, and not sufficient, for moral responsibility. Even if an agent has alternative possibilities, we would be reluctant to hold her morally responsible for her chosen action if this choice was made in an arbitrary, groundless way, or if it was the result of external manipulation, and so not really the agent's own. And the reason

clearly seems to be that the agent does not have an appropriate control over her own choice and action. So incompatibilists may insist on the necessity of alternative possibilities, but they must also hold that a control condition is required for moral responsibility, for otherwise the resulting freedom would amount to sheer arbitrariness, a clearly inadequate basis for moral responsibility. A reasonable incompatibilism, then, should defend—in Fischer and Ravizza's (1998) apt expression—"regulative control", that is, not merely the availability of alternative possibilities, but also control over which of them becomes actual, as the freedom-relevant condition for moral responsibility.

Incompatibilists, and especially libertarians, have defended a categorical interpretation of the alternative possibilities condition, against the conditional interpretation favoured by compatibilism. On their view, a merely conditional interpretation does not provide an adequate basis for moral responsibility. If the choice of an alternative possibility depends on a condition whose satisfaction is beyond the agent's choice, such as a difference in the past or in the natural laws, it cannot truly be said that choosing that alternative is within the agent's power. And so it is not justifiable to hold her morally responsible for her actual choice and action. On a categorical interpretation, an agent could do otherwise only if, keeping the past and the natural laws fixed, it was within her power to choose and follow an alternative course of action. Only if the agent satisfies the condition on this categorical reading thereof can it be justifiable to hold her morally responsible for her actual decision and action, since only then is it true, in the relevant sense, that she could have done otherwise. It is reasonable to think that, on this interpretation of the alternative possibilities condition, indeterminism is required for it to be satisfied.

As for the control condition, incompatibilists need not reject those aspects of this condition rightly emphasized by compatibilist analyses of it, such as volitional and intelligent or rational control over one's choices, as well as evaluative reflection on one's motives. On the contrary: if they want to put forward a reasonable proposal, they should accept (as some of them have done) that those aspects are constitutive of the sort of control required for moral responsibility. What they hold, however, is that they are not sufficient, and that the relevant control has to have an additional feature, namely ultimacy. Without this feature, agents could not justifiably be said truly to deserve praise or blame for their choices and actions, for they could not be their ultimate sources, and moral responsibility, in this deep sense of true desert, would not be possible. Since there cannot be either ultimate sources or ultimate control in a deterministic world, only an indeterministic world might be receptive to the satisfaction of this ultimacy requirement for moral responsibility. This is the hope of incompatibilists. And there is a rationale for this hope: only indeterminism would seem to provide room for fresh, absolute starting points, and so for ultimacy of control and source.

Though incompatibilists, especially libertarians, typically hold that ultimate regulative control (to use Fischer and Ravizza's useful terms again) is the freedom-relevant condition for moral responsibility, some incompatibilists may also base their case on a weaker condition. They are likely to be convinced by Frankfurt cases that alternative possibilities are not really required for moral responsibility. They will hold that all that is required is (in Fischer and Ravizza's terms) ultimate guidance control, a sort of control over one's decisions and actions that does not include alternative possibilities. Following recent usages, incompatibilists of this sort may be called "source incompatibilists". According to them, although alternatives are not necessary for moral responsibility,

incompatibilism still holds, for moral responsibility does require ultimate (guidance) control and this is not compatible with determinism. This contention is nowadays defended by Derk Pereboom, among others. (Pereboom also argues that control of this sort is ruled out by indeterminism as well. For this reason, he calls his own position "hard incompatibilism".)

The possibility of an incompatibilism that does not resort to alternative possibilities, but only to ultimate control, indicates a certain primacy of this latter condition over the former. This primacy can also be found in some distinguished libertarians, such as Robert Kane. Though he accepts the necessity of alternative possibilities for moral responsibility, he restricts the role of this condition as a basis for incompatibilism in favour of ultimacy of control and source: "Focusing on the power to do otherwise and alternative possibilities alone is just too thin a basis on which to rest the case for incompatibilism" (Kane 1996:59). In Kane's view, as we shall see, the importance of the alternative possibilities condition derives from its role in the ultimate control condition.

So even though indeterminism allows for alternative possibilities, this does not show that it allows for moral responsibility as well. Whether or not alternative possibilities are required for moral responsibility, ultimate control (of the "regulative" or merely "guidance" variety) is, according to incompatibilists, certainly required for it. In the preceding chapter we have seen that there are strong reasons for the truth of this incompatibilist necessity claim. Since, according to both incompatibilists and compatibilists, ultimate control is not possible if determinism holds, what libertarian incompatibilists should argue is that it *is* possible if *indeterminism* holds. Only then will they have made a convincing case for the possibility of moral responsibility.

Although indeterminism may seem to allow for fresh starting points, for ultimate sources or causes, what needs to be shown, in addition, is that these *ultimate* sources—if such there be—of a subject's actions are sufficiently under the subject's *control* for her to truly deserve praise or blame for performing them. Compatibilists have traditionally argued that indeterminism undermines control, for causally undetermined events occur by chance and events that occur by chance are not under anyone's control. Their idea seems to be that the only way an event can be controlled is by exercising control over its causes; it follows that an event that is not determined by causes cannot be controlled by anyone, nor can anyone be responsible for it. Indeterminism, then, erodes the basis of moral responsibility.

This traditional "chance" or "luck" objection against incompatibilism, which is usually known as the "Mind" argument, has been expressed in several versions. Alfred Ayer's is fairly representative of them all. In the middle of the last century, he argued against the libertarian ("the moralist", in his text) on this basis:

[T]he moralist…is anxious to show that men are capable of acting freely in order to infer that they can be morally responsible for what they do. But if it is a matter of pure chance that a man should act in one way rather than another, he may be free but can hardly be responsible…

To this it may be objected that we are not dealing fairly with the moralist. For…he does not wish to imply that it is purely a matter of chance that I act as I do. What he wishes to imply is that my actions are the result of my free choice: and it is because [of this] that I am held to be morally responsible for them.

> But now we must ask how it is that I come to make my choice. Either it is an accident that I choose to act as I do or it is not. If it is an accident, then it is merely a matter of chance that I did not choose otherwise; and [then]…it is surely irrational to hold me morally responsible for choosing as I did. But if it is not an accident…then presumably there is some causal explanation of my choice: and in that case we are led back to determinism.
>
> (Ayer 1954:17–18)

If, in order to prevent the choice from being arbitrary, the libertarian comes to hold that the choice itself is a result of the agent's character, the same dilemma arises, for we can then ask whether it is by accident that he came to form this character or not. Ayer's implicit advice to the libertarian is that, if she is really interested in defending the possibility of moral responsibility, she should seriously think of rejecting the view that moral responsibility is actually incompatible with determinism, and accept compatibilism instead.

Ayer's formulation of the chance objection may strike us, from a larger perspective on the development of the discussion about moral responsibility, as a bit naive. He is assuming that all causation is deterministic causation and that, if an event is not causally determined, then it occurs merely by accident or chance. Both assumptions can be, and have been, challenged by libertarians. The possibility that some causation may be probabilistic, and not deterministic, has been taken seriously. If this is right, Ayer's objection fails, for it is then possible to conceive of events that are caused, though not causally determined, and that thereby do not need to occur as a matter of pure chance. Choices could be among them. But it is not so easy to get rid of the background insight that underlies Ayer's objection. This insight can be given many formulations, and some of them can avoid the shortcomings of Ayer's version and deal successfully with libertarian replies to it. Let us now look at other possible ways of giving expression to that insight.

To begin with, the traditional worry about incompatibilist accounts of moral responsibility could be translated to our own terms, as a worry about the requirement of ultimate control. Ultimate control has two aspects: ultimacy and control. The ultimacy aspect points to the absence of sufficient antecedent causes, beyond the agent's reach, of the controlling factor, be it a practical judgement, a choice or even the agent herself, as in agent-causation libertarian theories. This aspect ensures that the agent can rightly be considered as the ultimate source of the action for which she is morally responsible. The control aspect, in turn, points mainly to the rational, volitional and evaluative factors emphasized in compatibilist analyses of freedom. This aspect is meant to ensure that the controlling factor does not arise by chance or mere luck. An arbitrary, chancy judgement or choice, even if it has no sufficient antecedent causes, is not an appropriate basis for moral responsibility. Both aspects have to be present for having such an appropriate basis. Now, the worry might be put by saying that the ultimacy aspect, which indeterminism is meant to make possible, undermines the control aspect, by introducing an element of sheer chance, luck or arbitrariness in incompatibilist accounts of moral responsibility. Alternatively, the worry might be expressed by saying that an agent cannot both be the *ultimate* source of her actions and keep ultimate *control* over them. One simply cannot have it both ways.

The point can be made more vivid by means of an example. This example is devised by Van Inwagen as an introduction to his version of the Consequence Argument against compatibilism, but it can also be used, somewhat paradoxically, to illustrate the worry we are dealing with, and so as part of a case against libertarian incompatibilism. Van Inwagen's example is the following:

> Let us suppose that there was once a judge who had only to raise his right hand at a certain time, T, to prevent the execution of a sentence of death upon a certain criminal, such a hand-raising being the sign, according to the conventions of the judge's country, of a granting of special clemency. Let us further suppose that the judge—call him "J"—refrained from raising his hand at T, and that this inaction resulted in the criminal's being put to death. We may also suppose that J was unbound, uninjured, and free from any paralysis of the limbs; that he decided not to raise his hand at T only after a suitable period of calm, rational, and relevant deliberation… I shall argue that, despite all these advantages, J could not have raised his hand at T if determinism is true.
>
> (Van Inwagen 1983:68–9)

One purpose of the example is to make readers think of the (absurd) consequences of determinism for our ability to do otherwise. Since it seems quite obvious that, at a certain moment in which we did not raise our hand, we could easily have raised it, then, if determinism implies that we could not, determinism is bound to be wrong. But the example can also be used to reflect about the consequences of indeterminism, which libertarians view as a requirement for ultimate regulative control and moral responsibility.

Suppose, then, that indeterminism is true, or at least that it holds where libertarians deem it especially important, namely at the moment prior to decision or choice, and go back to the example. On this assumption, J's decision at T not to raise his hand was undetermined. This means that there was some non-negligible probability that he would make a different decision instead, plausibly the decision to raise his hand, thereby granting clemency to the criminal. On the categorical interpretation of the alternative possibilities condition, defended by libertarians, this implies that a different decision, namely the decision to raise his hand, was compatible with exactly the same process and circumstances that preceded J's decision not to raise it. In this context, the actual decision may perhaps be said to be ultimately in J's hands, in that the chain of causes and effects came to an end before that decision, leaving it open that J decided as he did or in a different way. But now consider that, as things are in the actual process, J's decision to raise his hand would seem to be irrational and arbitrary. For, as Van Inwagen describes the process, J was perfectly sober and under no constraint, he was psychologically healthy, and his decision not to raise his hand came after "a suitable period of calm, rational, and relevant deliberation". We can assume, then, that he carefully weighed the reasons for and against granting clemency to the criminal and that he found the latter better than the former. In these circumstances, however, the libertarian insists that J could still have made the opposite decision. This decision would clearly be capricious and rationally inexplicable. But if the actual process of deliberation is compatible with it, then the links between J's reasons and his *actual* decision not to raise his hand are weakened.

This means that he loses rational control over the decision that he actually made, which so becomes more chancy and arbitrary as well. This is one sense in which indeterminism can be taken to undermine *control* over one's choices and actions.

The only way in which one could make rational sense of a decision of J's to raise his hand, and so to grant clemency to the criminal, would be to suppose that the process that would end with this alternative decision was in some respect different from the actual process. For instance, we can imagine that, in this alternative process, J is affected by a deep feeling of pity for the criminal, which leads him to assess the reasons he considers in a different way as well. This restores the links between reasons and decision, but it also involves a conditional interpretation of the alternative possibilities condition, which the libertarian will rightly consider insufficient to give agents *ultimate* control over their choices and actions. For in this case J's making an alternative decision at T will depend on a change in the past (the oncoming of the feeling of pity) which is no longer in J's power.

Another way of presenting the objection, which I borrow from A.Mele (cf. Mele 1999:98–9), adapting it to our case, is to imagine a close possible world, with the same past and natural laws, in which a twin of J's (call him J*) exists. The first difference between these two worlds arises only at T. At that moment, while J decides not to raise his hand, J* decides to raise it. On the assumption that J's decision is undetermined, this is clearly possible. But then, in Mele's words, "if…there is nothing about the agents' powers, capacities, states of mind, moral character, and the like that explains this difference in outcome, then the difference really is just a matter of luck" (Mele 1999:99). From this perspective, J's decision not to raise his hand appears to be a matter of luck or chance, and so not an appropriate target for moral praise or blame. Moral responsibility seems to lose its footing. The twin device is only another way of saying that, on the incompatibilist view, J's decision might simply have gone one way or another, and, if so, he is not truly praise- or blameworthy for it. The argument quickly generalizes. Moreover, replacing uncaused events by probabilistically caused ones does not seem to improve the libertarian's position.

Finally, let us see a powerful presentation of the general worry about incompatibilism that we are considering. This construal may be called the "Rolling-Back" argument (or the "Rolling-Back" version of the "Mind" argument). We can find it in a recent paper of Van Inwagen's (Van Inwagen 2000), in which he deals with the difficulties of his own libertarian position. For ease of exposition, we shall use Van Inwagen's own example. It features Alice, who, in a difficult situation, faces a choice between lying and telling the truth and freely chooses to tell the truth. Suppose now that free will is incompatible with determinism and that, therefore, Alice's telling the truth was undetermined. And then imagine that, after Alice's act, the world is reverted by God to the state in which it was, say, one minute before that act and then it is allowed to proceed "forward" again. The question is, what would have happened this second time. Would Alice have lied or would she have told the truth? Since her act was undetermined on both occasions, the only answer to this question is that she might have lied and also that she might have told the truth. Suppose that God causes this reversion one thousand times. What would have happened then? Van Inwagen goes on:

Well, again, we can't say what would have happened, but we can say what would *probably* have happened: sometimes Alice would have lied and sometimes she would have told the truth. As the number of "replays" increases, we observers shall—almost certainly—observe the ratio of the outcome "truth" to the outcome "lie" settling down to, converging on, some value… Let us imagine the simplest case: we observe that Alice tells the truth in about half the replays and lies in about half the replays…

(Van Inwagen 2000:14)

In this situation, the only reasonable option is, each time a new replay starts, to assign the same probability, about 0.5, to Alice's telling the truth and to her lying. But, as Van Inwagen writes,

[T]his surely means that, in the strictest sense imaginable, the outcome of the replay will be a matter of chance.

Now, obviously, what holds for [each] replay holds for all of them, including the one that wasn't strictly a *re*play, the initial sequence of events. But this result concerning the "initial replay", the "play", so to speak, should hold whether or not God bothers to produce any replays. And if He does not—well, that's just the actual situation. Therefore, an undetermined action is simply a matter of chance…

(Van Inwagen 2000:15)

Though Van Inwagen is directly concerned about how this argument threatens free will, its consequences for moral responsibility are even clearer. It would be totally unjustified to hold Alice morally responsible for her telling the truth, given that her telling the truth was a matter of chance. She might equally have lied. Though Van Inwagen deals with Alice's act of telling the truth, the argument obviously applies to her *choice* or *decision* to tell the truth as well. Again, indeterminism deprives agents of the control (and, a fortiori, the ultimate control) over their decisions and actions that would be required for them truly to deserve moral praise or blame for them.

This, however, is not the only problem faced by libertarianism. This approach has also frequently been accused of putting forward an obscurantist, anti-scientific metaphysics, by committing itself to entities and relations that can hardly be accommodated in a view of the world inspired by the natural and social sciences. This is the case in some agent-causation theories. Chisholm, for instance, postulated, in addition to causal relations between events, a special kind of causal relationship between agents and events that he called "immanent causation" (Chisholm 1964:28), but this sort of causation relation, quite unlike the ones that sciences seem to deal with, is also postulated by more recent theories as well. Equally distressful is Chisholm's conception of a free and responsible agent, clearly inspired by the requirement of ultimacy of source and control: "If we are responsible, and if what I have been trying to say is true, then we have a prerogative which some would attribute only to God: each of us, when we act, is a prime mover unmoved. In doing what we do, we cause certain events to happen, and nothing—or no one—causes us to cause those events to happen" (Chisholm 1964:32). Though not in such a frank and explicit form, views akin to these would seem to underlie other libertarian proposals as well. In this context, one can certainly understand P.F.Strawson's

dismissing reference to "the obscure and panicky metaphysics of libertarianism" (Strawson 1962:80). Though metaphysical obscurity is not, in itself, a decisive objection, a satisfactory version of libertarianism should indeed try to avoid being exposed to it.

Anyway, on the basis of the chance or luck objection, we now seem to have a general argument for SMR's premise C, namely: 1) Ultimate control is necessary for moral responsibility. 2) Indeterminism makes control, and a fortiori ultimate control, impossible. 3) Therefore, if indeterminism is true, moral responsibility is impossible. In the preceding chapter, we have argued for premise 1 of this argument. And this premise will not be denied by incompatibilists, anyway, since it is central to their view of moral responsibility. In the present section, we have seen arguments for premise 2. But, of course, it would be too rash to accept this premise on this sole basis. The arguments in its favour that we have gone through in this section are older and newer guises of the same basic suspicion that libertarianism has always aroused, namely that indeterminism would seem to introduce into the exercise of agency an element of chance or randomness incompatible with the agent's truly deserving praise or blame for her choices and actions. This suspicion is at least as old as Hume, who writes: "Where [actions] proceed not from some *cause* in the character and disposition of the person who performed them, they can neither redound to his honour, if good; nor infamy, if evil. The actions themselves may be blameable… But the person is not answerable for them…as they proceeded from nothing in him that is durable or constant" (Hume 1975:98). Of course, libertarians have been aware of this objection and have tried to argue that it is not actually justified. And, just as, in the compatibilist field, a significant difference in explanatory power and theoretical sophistication can be perceived between the initial, bold formulations of Hobbes's or Hume's and more recent approaches, such as Frankfurt's, Wolf's or Fischer's, a similar progress can also be noticed in recent libertarian theories of freedom and moral responsibility as compared with previous proposals. A central concern of these theories has been to counter the worries addressed against libertarianism that we have outlined here. The objections related to the role of luck and chance have received special attention, though the obscurity objection has also been taken seriously, and efforts have been made to counter it. In the next sections, we shall concentrate on an account that, in Alfred Mele's words, is "quite simply, the most thoughtful and detailed defence of libertarianism currently available" (Mele 1999:96), namely Robert Kane's admirable book *The Significance of Free Will*. We shall outline the basic structure of Kane's theory and try to see whether it can successfully dispel the worries that have been presented in this section.

To summarize, the challenge for libertarians is to show that agents can have ultimate control over their actions if indeterminism holds, and to show this without committing themselves to suspicious or mysterious entities and relations. These are, in fact, central theses of Kane's theory.

Kane's conception of free will

In his detailed and lucid libertarian approach to free will, Kane readily acknowledges the justice of several traditional objections to libertarianism. The charge of mystery and scientific obscurantism is seriously taken into account. Kane intends to develop his view without committing himself to entities and relations that are not also used in

non-libertarian approaches to free will and moral responsibility, so bringing libertarianism in line with the spirit and findings of the natural and social sciences and making it intelligible to scientifically formed minds. A libertarian view of free will need not contain more mysteries than those that are such for everybody, nor does it have to be less close to scientific thought than compatibilism or hard determinism. According to Kane, libertarians should seriously try to "show how an incompatibilist free will can have a place in the scientific picture of the world" (Kane 1996:213). In the same vein, far from rejecting those aspects of freedom and control that are compatible with determinism, Kane holds them to be worth defending and promoting, as they are an integral part of any reasonable view of free will. The ability to do what one wants to do, the power to endorse or reject one's motives reflectively and to assess them critically from the point of view of one's values, the capacity to form objectively correct values and to be appropriately sensitive and reactive to reasons: all these forms of control over one's motives and actions may arguably be necessary for free will and moral responsibility, and therefore highly valuable features of persons. In fact, rational and volitional control over one's choices and actions will be a central component of Kane's own concept of free will.

Again in line with compatibilism, Kane does not conceive of free will as a property or ability that is beyond the reach of natural and social causes. While some libertarian thinkers, such as Descartes or Sartre, conceived of free will as a transcendental property, immune to any influence by external causes, that is either perfectly and completely possessed by an agent or not at all, Kane insists that free will and moral responsibility can be had in different degrees, as they partially depend on various natural and social factors. So, against certain inhuman retributivist views of punishment that can derive from such absolutist approaches to free will, libertarians, according to Kane, can avail themselves of the same intuitions as compatibilists have about the sort of factors that mitigate an agent's moral responsibility.

Free will is not unaffected by particular social and political environments either. On the contrary, while totalitarian regimes are rather unwelcoming to its flourishing, socially and politically free settings provide favourable contexts for its development. Kane, then, rejects the frequently expressed view that social and political freedoms have nothing to do with the philosophical concept of free will. I think he is certainly right on this account. So, together with certain aspects of compatibilist views of freedom and control, Kane also values varieties of freedom, such as political and social freedom, that are plausibly held to be compatible with determinism. These are certainly aspects and varieties of freedom that we want to possess, and rightly so.

Kane insists, however, that there is an aspect of free will that we also crucially value and want to possess and that compatibilism cannot consistently afford. This aspect of free will is required for our deeds and achievements to have objective worth and for us truly to deserve praise or blame for them. It amounts to our capacity to be *ultimate sources* of those deeds and achievements (cf. Kane 1996:98). Nozick (1981) pointed to this aspect when he held that free will is valuable because it confers upon those beings who have it "originative value", that is, the capacity to introduce in the world radically new, unforeseeable value, not implicitly contained in the past history of the universe. In so far as ultimacy of source is required for objective worth and desert, it is required for moral responsibility, if this is understood in terms of true desert. This ultimacy condition is, for

Kane, the central feature of free will, which he defines as "the power of agents to be the ultimate creators (or originators) and sustainers of their own ends or purposes" (Kane 1996:4).

That agents can be such ultimate sources is only of interest for someone who values responsibility as involving true desert and objective worth. If one sees the practice of holding people (morally) responsible as only a useful, maybe indispensable, tool for controlling and modifying their behaviour to fit socially acceptable patterns, or if one sees that practice as just an expression of our natural tendency to display reactive attitudes, such as gratitude, indignation, remorse or resentment, addressed to other people or to ourselves, then compatibilist views of freedom and control will appear as all that is required for moral responsibility. But it seems to me that most of us are deeply convinced that, sometimes, our gratitude or indignation is grounded in the objective fact that its target is actually the ultimate origin of that which arouses those attitudes in us, so that she truly deserves them, with their associated praise or blame. I would think that this conviction is as natural in us as our disposition to react in such ways, so that, *pace* Strawson, abandoning this conviction would actually affect that disposition as well. We do not see our reactive attitudes as mere natural reactions, but as justified by facts about the way people choose and act.

Now it clearly seems that, if this conviction is to be true, then determinism must be false, for determinism does not leave any room for ultimate causes or causal origins. It is, then, ultimacy of source that makes compatibilist accounts of freedom and control seem insufficient to ground moral responsibility understood as true desert and gives rise to the incompatibilist claim that free will and moral responsibility cannot coexist with determinism.

Let us now see how Kane construes the ultimacy of source condition for free will, which is itself required for moral responsibility understood as true desert. He does so in terms of what he calls "Ultimate Responsibility" (UR). Let us recall Kane's characterization of this necessary condition of free will:

(UR) An agent is *ultimately responsible* for some (event or state) E's occurring only if (R) the agent is personally responsible for E's occurring in a sense which entails that something the agent voluntarily (or willingly) did or omitted, and for which the agent could have voluntarily done otherwise, either was, or causally contributed to, E's occurrence and made a difference to whether or not E occurred; and (U) for every X and Y (where X and Y represent occurrences of events and/or states), if the agent is personally responsible for X, and if Y is an *arche* (or sufficient ground or cause or explanation) for X, then the agent must also be personally responsible for Y.

(Kane 1996:35)

As an initial step to a proper understanding of this condition, let us comment on this text as follows. If an event is or issues directly from a voluntary act or choice of the agent, with respect to which she could have done otherwise, then she is already personally responsible for that event and can be ultimately responsible for it. But if the corresponding event is not a voluntary act itself or issues directly from causes other than a voluntary act of the agent, then, for the agent to be ultimately responsible for this event, these causes must themselves be causally traceable to voluntary acts of hers, which she is

personally responsible for and regarding which she could have done otherwise. So, at the root of any event for which an agent can be ultimately responsible, there has to be an event for which the agent is personally responsible, in the sense that this event either is or directly issues from a voluntary act or choice of hers with available alternatives. Suppose, for example, that, at a certain time, I intentionally perform a particular action with some moral import, and that, given my present character, values and preferences, my doing otherwise would be inexplicable. Then, in order to be ultimately responsible for that action, I must be personally responsible for the character, values, and preferences it flows from; and this, in turn, requires that I have made a significant causal contribution to my presently having that character, values, and preferences through some voluntary past choices and actions, with regard to which I could have done otherwise.

Let us now go a bit further. The expressions "voluntarily" or "willingly", which appear in the statement of the UR condition, are technical expressions. They include a reference to reasons. To act voluntarily or willingly is to do what one wills to do, for the reasons one wills to do it, without coercion or compulsion, whereas willing to do something, in turn, is having reasons or motives one wants to act on more than one wants to act on other reasons (cf. Kane 1996:30). The notion of will in play here is Kane's notion of rational will, which he understands as "a set of conceptually interrelated powers or capacities" connected with practical reasoning and decision-making (cf. Kane 1996:22). It can be seen, then, that UR includes the idea, central in compatibilist analyses of free will and moral responsibility, of having volitional and rational control over one's choices and actions. What UR adds to this idea is the requirement that this volitional and rational control be ultimate. This means that the causal history of a certain choice or action cannot include sufficient causes over which the agent does not have volitional and rational control. In other words, the factors that explain a certain choice or action of a particular agent at a particular time must themselves be causally produced by voluntary acts, in the sense indicated, of this agent. From this point of view, Kane's Ultimate Responsibility can rightly be seen to be a version of the Ultimate Control condition for moral responsibility that we have been discussing in this and the preceding chapter.

The UR condition, as Kane himself acknowledges, seems to lead to a vicious regress of choices or actions. This regress can only be stopped by actions or choices that themselves have no sufficient causes. Thus, ultimate responsibility, and so free will and moral responsibility, can only exist if determinism is false. But Kane's task is to show how these regress-stopping actions, which are required for an agent to be ultimately responsible for any action at all, and so required for her free will and moral responsibility, can be voluntary acts themselves and remain under the agent's rational and volitional control. These regress-stopping actions, then, play an absolutely crucial role in Kane's theory of free will and moral responsibility. Kane calls them "Self-Forming Actions (SFAs)" (cf. Kane 1996:75). The agent must be responsible for these actions in a direct way, by having chosen them voluntarily from (categorically) available alternatives, not by means of their relationship to earlier actions, if she is to be ultimately responsible for any action at all. SFAs, then, cannot have a sufficient cause or explanation if they are to stop the regress that Kane's Ultimate Responsibility initiates.

From a structural point of view, in Kane's theory of free will and moral responsibility SFAs play a foundational, regress-stopping role that closely parallels the role that basic actions play in some theories of action, or that certain kinds of basic, non-inferential

beliefs play in foundational theories of knowledge. So, in the same way in which, according to those theories of action, there could be no actions at all unless there were basic actions, and, according to those theories of knowledge, no belief could be justified at all unless there were beliefs that are not justified by other beliefs, in terms of Kane's theory there could be no actions for which an agent is ultimately responsible, and so morally responsible, unless there are voluntary actions for which the agent is directly personally responsible, that is, self-forming actions. These actions must be undetermined, that is, they must have no sufficient causal explanation, for otherwise they could not be ultimate causes; the agent's pre-existing character and motives must not determine them; rather, one of their functions is precisely to form that character and those motives, so that, if later actions are explained by the agent's character and motives, the agent can still be said to be ultimately responsible for those actions.

As we have already pointed out in several places, the role of the alternative possibilities condition in Kane's theory is to be understood in the context of the ultimate control condition. The freedom-relevant condition of moral responsibility is just ultimate control, understood as UR. However, alternative possibilities must be present at some points in an agent's life history for ultimate control itself to be possible. Some actions have to be such that, with regard to them, the agent could have done otherwise, in a *categorical* sense of this expression. So, for free will, understood in terms of ultimate responsibility, to be possible, "some free actions must be undetermined. They must be capable of occurring or not occurring, given *exactly the same past and laws of nature*" (Kane 1996:106). Not all actions for which an agent is ultimately responsible have to satisfy this condition, categorically understood. It may well be that, at a certain moment, given the agent's character, values and motives, her will is clearly settled in favour of an option, so that her choosing in a different way would be inconceivable. But if the agent is to be ultimately responsible for this choice, it must be traceable to earlier choices and actions with respect to which her will was not settled and she could, categorically, have done otherwise. Whether or not categorical alternative possibilities are present in other actions, they *must* be present in SFAs if ultimate responsibility, and so moral responsibility, is to be possible.

Now, since SFAs are the foundations of moral responsibility, understood as true desert, so that the responsibility an agent may bear for any action of hers has its last roots in them, the agent herself must be responsible for her SFAs. Otherwise ultimate responsibility would be a miracle. And the traditional objection to incompatibilism, namely that indeterminism seems to undermine responsibility and control, arises again with regard to SFAs. As we have said, these basic, foundational acts must have no sufficient cause or explanation, for otherwise they could not stop the regress that threatens moral responsibility. Not even our character, preferences, and motives should be allowed to provide such a sufficient cause or explanation. But then SFAs would appear to be chancy, arbitrarily produced events, which, it seems, nobody could be responsible for (cf. Kane 1996:37).

This problem is, of course, a version of the "chance" objection to incompatibilism that we were looking at in the preceding section. Let us see how Kane proposes to deal with this vexed question.

The most general form of this objection is to hold that undetermined actions (or events generally) are chancy and arbitrary, and therefore not such that agents can have control over and be responsible for them. Kane thinks that, in this general form, the objection can be met. Following Anscombe, he avails himself of the notion of probabilistic or non-deterministic causation. In terms of this notion, an event can rightly be said to cause another even if the latter is not unavoidable given the former. In connection with this, an agent can also rightly be said to have produced a certain outcome and to be fully responsible for it even if this outcome was only probable given what she did. An example of Kane's vividly illustrates this point. The example features a nuclear facility employee who, "with evil intent", puts a piece of radioactive material into the desk of an executive whom he hates. Now,

> [w]hether the executive actually gets cancer may be genuinely a matter of chance, since that outcome depends on the occurrences of undetermined quantum radiations and mutations… But if the executive gets cancer, there is no question that the employee should be held responsible.
>
> (Kane 1996:55)

What this and similar examples show is that responsibility can exist in the absence of causal determination. So, the mere fact that SFAs are undetermined does not imply that an agent is not responsible for them. Probabilistic causation may suffice for responsibility, given other appropriate conditions.

Though Kane holds that this move answers the general form of the chance objection, it seems doubtful to us that it really does. Kane's response seems to involve a conflation between actions and outcomes of actions. Suppose that the executive gets cancer as an outcome of the employee's putting the radioactive material in the drawer. Though this outcome is genuinely undetermined, and not under the agent's control, the agent is rightly held responsible for it. But what still has to be shown is that he would also be responsible for that outcome if his action were equally undetermined and not under his control. Now, SFAs are not outcomes of actions, but actions (or choices) themselves. If they are outcomes, they are probabilistic outcomes of non-actional factors, such as motives or reasons. But if the control one agent has over her actions or choices, given her motives and reasons, is of the same kind as the employee has over the executive's getting cancer, given his action, it is doubtful that she can truly be responsible for those actions (and therefore for their outcomes). If the executive's getting cancer is a genuinely random outcome, it is not under anyone's full control. But this would also be the case with actions if they were related to an agent's character and motives in the way in which the executive's getting cancer is related to the employee's action. Obviously, appealing to probabilistic causation still leaves the chance problem, and the related problem of control, even in its more general form, open.

This difficulty, whether or not Kane is aware of it, is actually related to a deeper form of the chance objection, which, as Kane himself acknowledges, the appeal to probabilistic causation is not sufficient to solve. This form has to do with the categorical version of the alternative possibilities condition, according to which it must be true, for any SFA, that it can occur or not, given exactly the same past and natural laws.

A good illustration of this problem is Van Inwagen's example of the judge. We saw in the preceding section that, given the circumstances of the case, including the actual process of deliberation, the judge's making an alternative decision would be arbitrary and irrational. So, even if we allow that, given indeterminism, the judge's decision was not inevitable, it certainly seems that, unlike an alternative decision, it was rationally made, in the light of carefully considered reasons and a calm process of deliberation. Given these very same reasons and deliberation process, deciding otherwise, though causally possible, would have been arbitrary and capricious. Even if the judge might be said to be the ultimate source of this alternative decision, it would have been made without rational grounds, and so without appropriate rational control. Ultimacy and control appear to conflict with one another. And, in so far as indeterminism makes this sort of irrational and capricious alternatives possible, it thereby undermines an agent's rational control over her own choices and so threatens moral responsibility. We do not want moral responsibility to rest on arbitrary choices. Ultimacy of source must go together with rational control if it is to ground moral responsibility. And the problem is how to ensure this if indeterminism holds.

Let us point out that Kane's approach can deal with the judge's and other similar examples, that is, with cases in which, given the agent's character, reasons and values, one option is clearly superior to its alternatives, for Kane does not require, of any choice and action which an agent is ultimately responsible for, that the agent could, categorically, have done otherwise with respect to *it*. But if particular choices or actions are such that the agent could not, at least with voluntary and rational control, have done otherwise, then, in order for the agent to be ultimately responsible for them, they have to be appropriately traceable to choices or actions of which it is true that the agent could, categorically, and with voluntary and rational control, have done otherwise. These are the self-forming actions (or choices). For these choices or actions, then, it has to be the case that they are rationally and voluntarily performed, and that, had the agent chosen or acted otherwise, these alternatives would also have been under the agent's rational and voluntary control.

This deeper version, in terms of arbitrariness or irrationality, of the chance objection to incompatibilism is what Kane calls "the problem of plurality". This is the problem of explaining how an agent who faces a choice between (at least) two options can "choose *either* option rationally, voluntarily, and under voluntary control, given the same past and laws of nature" (Kane 1996:115). Either-way rationality, voluntariness and voluntary control are what Kane calls "the plurality conditions". Again, not all actions and choices for which an agent is ultimately responsible have to meet these conditions. But at least SFAs have to satisfy them if UR, and so moral responsibility, for *any* action or choice is to be possible.

Now, the whole issue of the possibility of moral responsibility, given indeterminism, is concentrated in the question whether SFAs can solve the plurality problem.

Let us have a closer look at the plurality conditions. They certainly look very demanding. Imagine an agent who confronts a choice between A and B. The plural rationality condition requires that an agent who confronts this choice and who, after deliberation, rationally chooses A, could, equally rationally, have chosen B instead, given exactly the same past, and especially the same process of deliberation through which she came to choose A. This looks rather strange. For, if the agent concluded that, all things

considered, A was her best option, her choosing B, in those same circumstances, would appear not to be a rational choice. Plural voluntariness requires that, if this agent chose A, and the choice was voluntary, she could, also voluntarily, have chosen B instead. But remember that acting voluntarily, or in accordance with one's will, is to do what one wills to do, for the reasons one wants to do it, in the absence of coercion or compulsion. And willing to do something is to have reasons or motives one wants to act on more than one wants to act on any other reasons (for doing otherwise). Applying this to our agent's choice, if she voluntarily chose A, this means that she willed to choose A, which in turn implies that she had reasons for choosing A which the agent wanted to act on more than the reasons for choosing B. But then it seems that her choosing B could not possibly be voluntary. Finally, plural voluntary control relates to the will's control over actions rather than choices. It is the ability "to *bring about* any one of the options…*at will* or *voluntarily* at the time" (Kane 1996:111), or, as Kane also puts it, it is the ability "to do *whatever you will* (or most want) to do, *whenever you will to do it,* for the reasons you will to do it" (Kane 1996:111). This condition inherits the problems it has from the other two.

These conditions look incredibly strong. Why insist on plural rational and voluntary control instead of simply requiring that the option one actually chooses be rational and voluntarily controlled, no matter whether an alternative option should be so as well? The reason is that these conditions (which Kane labels "one-way" rational and voluntary control) would be too weak to support UR. If an agent's action or choice is one-way rational and voluntary, this means that the agent's will is already settled by virtue of certain motives or reasons that she has. But then her choice or action will be explained by her having those motives or reasons, and it may be that she is not responsible for having them; and, if she is not, she cannot be ultimately responsible for her choice or action either. As Kane rightly points out, "*the plurality conditions are entailed by UR* for undetermined actions" (Kane 1996:112). In terms more familiar to us, what is required for ultimate responsibility and so for moral responsibility, in the sense of true desert, is ultimate regulative control. In Kane's theory, the events that are meant to satisfy these plurality conditions, or ultimate regulative control, are the SFAs, the regress-stopping actions required by UR. Not all choices or actions that agents are responsible for need to meet the plurality conditions. Agents can face some of these choices or actions with decisive reasons and with a will that is definitely settled in favour of one option: think of what we called "Luther cases". However, if all choices or actions were of this kind, agents could not be ultimately responsible for anything they choose or do, for they could not be said to have formed the wills and characters on which they act (cf. Kane 1996:114). There have to be free and responsible choices or actions that are plurally rational and voluntary, by performing which agents build up their own wills and characters. These are the SFAs.

How could SFAs satisfy the plurality conditions, then? It would be tempting, but wrong, to think of SFAs as a result of cases in which, at the end of her deliberation, the agent has equal reasons for going one way or another. It would be wrong because in these cases ("Buridan" or "liberty of indifference" cases, to use traditional labels), as Kane himself points out, "instead of one choice (A or B) being arbitrary relative to the prior deliberation, both would be arbitrary" (Kane 1996:109). There is, however, a connection between liberty of indifference and the intuition behind Kane's SFAs, for, in some sense,

it must be true of them that the agent will be torn between conflicting sets of reasons or motives, neither of which will appear to her as a priori stronger than the other, if these acts are to be "will-setting" and "regress-stopping". What, then, do SFAs consist in if they are not Buridan cases?

To respond, it is crucial to focus on an important subset of SFAs, namely SFWs, or self-forming willings. They include: moral choices, prudential choices, efforts of will sustaining purposes, attention efforts directed at self-control and self-modification, practical judgements and choices, and changes of intention in action. Let us focus on the first two. In moral and prudential choices there is a conflict between what an agent thinks she should or ought to do and her actual wants or desires. In these cases, unlike Buridan cases, it is not that the agent is indifferent between *equal* reasons to go one way or another; rather, what the agent faces is *incommensurable* sets of reasons in favour of alternative choices. In Kane's theory, this notion of incommensurability replaces the traditional idea of indifference. Paradigmatic cases of exercising free will are not Buridan cases, but cases in which agents are "torn between conflicting internal points of view that represent *different and incommensurable visions of what they want in life* or what they want to become" (Kane 1996:199). This is why Kane calls regress-stopping actions "self-forming": they shape the agent's self, her character, motives and will, so that she can be said to be ultimately responsible for her own character and motives and, thus, for further actions that are explained by these factors.

Essential to the idea of an SFW is the notion of an "effort of will" by the agent which she exercises between her prior reasons and motives and her choice. In the case of moral choices, where moral considerations conflict with self-interested motives, this effort of will is required to "overcome temptation" and make moral reasons prevail. And something of the sort happens in prudential choices, where desires for an immediate good conflict with long-term benefits.

These efforts of will that take place between reasons or motives and choice are crucially *indeterminate,* which makes the choice *undetermined* (cf. Kane 1996:128). There is, then, a causal gap between reasons and choices that is both produced and occupied by the indeterminacy of the agent's efforts of will. We can now see how SFWs and SFAs differ from merely random events that the agent cannot control, as happens, in the example of the nuclear facility employee, with the executive's getting cancer. Whereas, in this case, there is nothing that the agent can do after putting the radioactive material in the drawer except wait and see, there is actually something that an agent can contribute, in SFWs, between her prior character and motives and her choices, namely her effort of will. Kane's approach, then, can cope with the difficulty we raised above, according to which actions would be random and not under the agent's control if they were related to her character and motives in the way in which the executive's getting cancer is related to the employee's action. It is the agent's effort of will that makes it possible to tell these two relationships apart.

In accordance with his purpose of developing a libertarian theory that is not at odds with scientific thinking, Kane tries to explain the indeterminate character of efforts of will in SFWs, and the resulting undetermined character of the choice, with a hypothesis about the nature of neural processes that take place in the agent's brain when such an effort of will is being made. Following the suggestions of several scientists and philosophers, he supposes that these neural processes are chaotic processes, in that tiny

changes in their initial conditions give rise to big differences in the final result. And to this picture he adds quantum indeterminacy in the subatomic particles constituting the neurons, so that, at a macro level, the chaotic nature of the brain system amplifies indeterministic changes at the quantum level. Suppose, finally, that neural processes function as non-equilibrium thermodynamic phenomena, in that, in certain conditions, they reach unstable bifurcation points at which the systems may evolve in different ways, depending on those undetermined quantum changes at the micro level. Kane also refers to the effects of chaotic amplification on neural networks, in the sense of connectionist models of the brain, and to the capacity of these networks for probabilistic self-organization after changes in the connection weights (firing potentials) of individual nodes in the network.

In the case of moral and prudential choices, two neural processes struggle to reach an activation threshold and this is phenomenologically experienced by the agent as an inner conflict between contrary sets of motives within her will. Though the prior character and motives explain the agent's effort to resist temptation, they, together with the effort, do not provide a sufficient explanation of the final choice. This is genuinely undetermined in that the effort itself is indeterminate. Given the indeterministic nature of the whole process, the existing motives and character influence, but do not determine, the final outcome. Now, "the indeterministic chaotic process is also, experientially considered, the agent's effort of will; and the undetermined outcome…[is] the agent's choice" (Kane 1996:147). Of course, the old problem of the relationship of mind and body is present in this picture, but, as Kane insists, this is a difficult question for everybody, and not only for libertarians.

Let us see, in this context, whether SFWs satisfy the plurality conditions. Kane holds that they do. As for plural rationality, it seems that the agent will have reasons for her choice, whatever it finally is, and that this choice will have been (probabilistically) caused by those reasons. As Kane writes, "for SFWs, each outcome is rational for different and incommensurable reasons" (1996:178). Plural voluntariness, however, would seem to be more problematic. For, in order to satisfy this condition, the agent must will the choice, whatever it is, and to will the choice is to have reasons or motives for choosing the way that the agent wants to act on more than on the reasons for the alternative choice. Now, imagine a conflict between moral and non-moral reasons as a result of which the agent finally chooses for moral reasons. In order for this choice to have been willed, and so voluntary, it seems that the agent will have wanted to act on the moral reasons more than on the non-moral ones. But then, had she chosen for her non-moral reasons, this choice would not have been voluntary, and plural voluntariness would not be satisfied. This, in turn, would retrospectively affect plural rationality, for the alternative choice would not have been fully rational: though the agent had reasons for it, she wanted to act on other reasons more than on these ones.

Kane's proposal for solving this important problem is to deny that, prior to her choice, the moral (non-moral) reasons were such that the agent wanted to act on them more than on the non-moral (moral) ones. Her will was unsettled. She wanted to act on one set of reasons and also wanted to act on the other set. Now it is her actually *choosing* for one set of reasons that makes that set of reasons the one she wants to act on more: "…The agents will *make* one set of reasons or motives prevail over the others then and there *by deciding*…[B]oth options are wanted and the agents will settle the issue of which is

wanted *more* by deciding" (Kane 1996:133). So, whichever way the choice goes, it will satisfy plural rationality and voluntariness.

This proposal is also the clue for answering the question of how SFWs can satisfy plural voluntary control. Though, given that the effort of will is indeterminate and the choice undetermined, the agent cannot have *prior* or *antecedent* voluntary control over her choice, she can none the less have voluntary control over it *when* she actually chooses. As Kane writes: "It does not follow that because you cannot guarantee which of a set of outcomes occurs beforehand, you do not control which of them *occurs, when* it occurs" (Kane 1996:134). Agents can have antecedent, one-way voluntary control on many occasions, but this presupposes that their wills are already settled. Now the function of SFWs is to make agents ultimately responsible not only for their actions but also for their wills, and so, in situations of SFWs, the agents' will must not already be settled. It is their choosing itself that settles their wills, and contributes to forming their character and motives.

So, if SFWs satisfy these demanding conditions, and agents can perform them, they can be said to have ultimate rational and voluntary control over (at least some of) their choices and actions, and so they can be morally responsible for them, in the deep sense of being truly praise- and blameworthy for (at least some of) their deeds.

In the light of all this, it seems that there is an important Aristotelian theme in Kane's theory. Free will and moral responsibility are, in an important sense, a question of developing appropriate habits and dispositions, and so of building up one's own self and character. In many, maybe most, cases of deliberation and decision-making in which the agents' will is already settled, they will be responsible for those decisions, and the corresponding actions, in an indirect way, namely by being responsible for having become the sort of persons that find certain reasons more attractive than others and are moved to act on them more than on other reasons. And they will be responsible for having become so by virtue of SFAs and SFWs that they performed in their past life history. As Kane points out, moral responsibility not only concerns particular actions but also the construction of one's character and habits, the development of "virtuous and vicious dispositions through one's own efforts, choices, and actions" (Kane 1996:180, 181). Aristotelian and virtue ethics themes resound, then, with some clarity in Kane's approach to moral responsibility.

Finally, in Kane's criticism of Kantian moral rationalism we can also find a curious return to some aspects of Hume's moral psychology. More precisely, Kane would seem to embrace the Humean idea that reason alone is motivationally inert, so that being moved to act necessarily involves a desire. Concerning moral choices, he denies, against Kant, that they are rightly conceived as motivated by reason alone. Correspondingly, immoral choices are not correctly seen as determined only by desire either. As he writes, "free will…involves reason-*plus*-desires on one side versus reason-*plus*-desires (of different kinds) on the other side… Reason may tell you that something is your duty, but you will not do it unless you also desire to do your duty" (Kane 1996:207). The Humean flavour of this remark is unmistakable.

This completes our exposition of Kane's theory of free will and moral responsibility. In the next section, we shall attempt to evaluate this proposal.

Difficulties for Kane's libertarianism

Let us see how Kane's theory fares with regard to the so-called "Mind" objection to incompatibilism, according to which indeterminism turns an agent's choices and actions into a matter of luck, chance or accident. In the first section of this chapter we have presented several versions of this objection, and in the second section we have seen how Kane could deal with some of them. Against Ayer's version, Kane resorts to probabilistic causation in order to show that an agent can be rightly held to be morally responsible for an outcome even if it is only probable, given the subject's action. Kane tries to show this by means of examples, such as that of the nuclear facility employee. We pointed out, however, that examples of this sort do not go far enough in the way of answering the objection. They focus on the relationship between a subject's action and an (undetermined) outcome of it. Once the subject has done her job, there is nothing else for her to do than passively wait and see what happens. Whether a certain outcome finally occurs (e.g., whether the nuclear facility executive gets cancer) is beyond her control. We pointed out that, if the agent's motives and reasons related to her choice and action in this kind of way, she could not be said to control her choice and action either. There is, however, in this case, and especially in SFW situations, an effort of will that allows the agent to intervene actively in the process between her motives and her choice, instead of passively waiting for it to develop. Even if, given the agent's reasons and effort of will, the choice is still undetermined, this in itself does not imply that it is purely chance and beyond the agent's control. Now, whether this move can finally meet Ayer's version of the "Mind" objection depends, then, on whether SFWs, of which efforts of will are a central component, are rightly said to be under the agent's rational and volitional control. It is important, therefore, to address this issue by carrying out a detailed examination of SFWs.

At the end of the preceding section, we indicated the presence of a Humean element in Kane's view of motivation. Let us now point out that Kane's view of the psychological structure of agents in SFW situations would seem to be pre-Frankfurtian and pre-Watsonian, and so, in a sense, also Humean, in so far as Frankfurt and Watson intended to overcome Hume's psychological picture of choice, which they rightly thought of as too simple to account for free will. In Kane's SFW situations, in fact, agents are torn between two incommensurable sets of reasons, each of which contains both rational considerations and desires. They want to act on one set and they also want to act on the other but, prior to their choice, they cannot be said to want to act on one of those sets more than they want to act on the other. This means that they do not have second-order desires or preferences about which of the two sets they want to be effective in leading them to act, or a system of values that ranks acting in one way higher than acting in the other. This is so because, if they had such second-order desires or value systems, then, prior to their choice, they would want to act on one set of reasons more than on the other, or they would judge one set to be better than the other, and then acting on the latter would not be rational or voluntary, in Kane's sense. In other words, if, in SFWs, agents were to have such a fixed value system or second-order attitudes before choosing, SFWs would not satisfy the plurality conditions and could not be regress-stopping.

Now, Frankfurtian second-order desires and Watsonian valuational systems play the role of a standard on which options can be assessed previously to choice, so that the agent can be said to decide and act with or without (or with more or less) rational and volitional control, depending on whether or not her decision and action are actually caused by those second-order desires or value systems or, in other terms, on whether or not her decision and action meet the standard established by them. In Kane's theory, instead, it is SFWs and SFAs themselves that are supposed to set up the standard, in that it is by performing those basic choices and actions that agents build up their own wills, the motives on which they most want to act, their values and the second-order volitions that arise out of them. In fact, the point of calling these basic choices and acts "self-forming" is that their central role is to form and shape the agent's self and character. Thanks to this, when the agents' wills are settled, and there is, prior to decision, a motive on which they want to act more than on other motives, they can be said to be the true and ultimate sources of the ensuing decisions and actions, and so to be morally responsible for them, in the sense of truly deserving praise or blame.

There is, then, a crucial reason for the pre-Frankfurtian and pre-Watsonian psychological picture of SFW situations in Kane's libertarian theory. This picture is intended to ensure that agents have ultimate (rational and voluntary) control over their own decisions and actions, in that they are the ultimate sources of the ends, purposes and motives on which they decide and act. This picture, then, is an essential precondition of ultimate control. If we allow agents to face *any* choice situations equipped with already-formed reflective standards, with values or second-order volitions that incline their wills in one particular direction, by making them want to act on some motives more than on others, or judge some motives to be better than others, then no choice, whichever way it goes, will satisfy plural rationality and voluntariness, and then ultimate responsibility, and so moral responsibility, will appear to be impossible.

At this point, we can already perceive some aspects of the dialectic involved in the question of free will and moral responsibility that make this problem or set of problems appear so tantalizing.

As we saw, compatibilism, even in sophisticated versions such as Frankfurt's or Watson's (or, for that matter, Wolf's or Fischer's), has a seemingly insoluble problem in ultimacy of source or origin, which is why it tries to reject the necessity of this condition for moral responsibility. However, it enjoys a corresponding advantage for what regards rational and volitional control. In providing agents with reflective standards, prior to choice, for assessing reasons, it gives a clear sense to the idea of these agents having rational and volitional control over their decisions. In Watson's view, for example, they have this control just in case their wills conform to their values; and corresponding accounts of control derive from the other versions.

In Kane's approach, however, the situation would seem to be exactly the opposite. It can give a clear sense to the idea of ultimacy of source, and set out the conditions that should be met for the agents' being such ultimate sources, which include indeterminate efforts of will and undetermined SFWs, but, in denying agents in SFW situations comprehensive reflective standards, prior to choice, by which to measure and assess the several sets of reasons and the corresponding options, it stumbles on the problem of their rational and volitional control over their choices.

Kane insists that SFW situations are not rightly conceived as Buridan cases, in which agents face a choice between, so to speak, equal amounts of the same good. Kane does not want SFW situations to be of this sort, for otherwise choice in such situations would be arbitrary. But this reason does not seem to be right. In fact, for rational agents, these cases are quite easily solved, for both options are commensurable: they are on the same scale. The only problem is that they reach the same mark on the scale. But then, tossing a coin, or something equivalent to it, is the right and rational decision in cases like this. And this decision is far from arbitrary. In Kane's SFW situations, however, options, and reasons for each of them, are not on the same scale: they are incommensurable. But then, in the absence of a rule or standard, prior to decision, for guiding the assessment, choosing between them becomes an arbitrary issue, over which the agent cannot have rational and voluntary control. Imagine, for example, a moral choice, in which the agent has two sets of incommensurable reasons, moral and non-moral, in favour of each of two options. In this case, having a second-order volition to choose and act for moral reasons, or a valuational system that ranks moral reasons higher than non-moral ones, provides a rule or standard that can guide the choice between the options, in that it settles the question of which set of reasons the agent values, or wants to act on, more. However, in the absence of such second-order attitudes or value systems, or something equivalent to them, there is no such standard or rule, and so no saying which of those two sets the agent values, or wants to act on, more. So the choice is bound to be made without rational and voluntary control.

Kane would deny this. According to him, far from it being the case that neither choice is under the agent's rational and voluntary control, he will insist that either choice is. As he writes, "for SFWs, each outcome is rational for different and incommensurable reasons" (Kane 1996:178). And each outcome is also voluntary in that the agent wills each of them. This means that, in choosing for either, the agent wants to act on the reasons for it more than on the reasons for the opposite. This, however, cannot be the case before the choice is made. As we have seen, Kane agrees that, in SFWs, the agent does not have *antecedent* voluntary and rational control over the choice, but he insists that the agent enjoys the required control at the very moment of choosing (cf. Kane 1996:115). It is not that, prior to decision, the agents want to act on one set of reasons more than on the other; but, whichever the set of reasons which they choose, these reasons will be the ones they want to act on more, and so their choices will be made with rational and voluntary control, for the agents "will have made those reasons the ones they wanted to act on more than any others *by* choosing for them" (Kane 1996:135).

But this move cuts no ice. It guarantees that, in SFW situations, whichever the reasons the agent chooses and acts on, her choice will be made with rational and voluntary control by definitionally linking "reasons that the agent wants to choose and act on more" with "reasons that the agent actually chooses and acts on". To say on this basis that, in SFW situations, either choice will be rational and voluntary is as empty as the traditional contention that an agent always acts for her stronger reasons, when which the stronger reasons are is made to depend on what reasons she in fact acts on. Kane's move distorts the very notion of a reason for choosing and acting, as well as the related notion of a rational choice and action. A rational choice is made on the basis of the reasons, among those that the agent considers, that she *finds or perceives as* better, stronger or more convincing; so it would seem to be constitutive of a reason for choice or action that the

strength or the conviction force it may carry does not depend upon the agents' will. But just the opposite is precisely what happens in SFW situations according to Kane, for remember that in these situations "the agents will *make* one set of reasons or motives prevail over the others then and there *by deciding*" (Kane 1996:133). In SFW situations, prior to decision, agents cannot find one set of reasons stronger or more convincing than the other, either in itself or as judged from a reflective standard, for, if they do, SFWs will not satisfy plural rationality and voluntariness, and so will not be regress-stopping. And the space left by this prior perceived conviction force or by such a standard is occupied by sheer, baseless *decision*. Kane's libertarianism, then, would seem to involve a sort of sheer *voluntarism* or *decisionism*. It would side with Plato's Euthyphro in holding that something is good because the gods want it, instead of the gods' wanting it because it is good, or with medieval views of God as essentially constituted by Will rather than Reason. In short, the cost of *ultimacy of source,* in Kane's theory, is loss of *rational control*. This old problem of libertarianism surfaces again in this sophisticated approach.

Kane's likely reply might be that a truly arbitrary choice would be one that is made in the absence of any reasons at all. And this is not the case with SFWs. In SFW situations the agent has (different and incommensurable) reasons for either choice. We must agree with this. Kane does not conceive of agents in SFW situations as *tabulae rasae*. They face these situations equipped with sets of reasons and motives, and with a certain character. These factors, however, he would insist, do not causally determine the choice, although they probabilistically cause it. This probabilistic causal relation is important in order for the choice to be connected with reasons, and so for it not to be arbitrary. Even if these reasons do not determine the decision, they certainly influence it, as Kane repeatedly contends. But then, not any decision may arise out of any conglomerate of reasons, where by "conglomerate of reasons" I understand the sum of the sets of reasons in favour of either choice or outcome. The particular conglomerate of reasons with which an agent faces an SFW situation limits the range of possible (rational) outcomes. But if so, then for the actual choice to have its ultimate source in the agent, the question about the origin of such a conglomerate is important. If the choice is to have its ultimate origin in the agent, the conglomerate that makes it possible has to be the effect of even earlier SFWs, with their own conglomerate of reasons, with regard to which the same question can be asked again. This would not be a regress of sufficient causes or motives, since it is agreed that the conglomerate does not causally determine the decision, but it would still be a regress of enabling or necessary causes of the (rational) choice in a particular SFW situation. This regress could only be stopped by an initial SFW that did not arise from any conglomerate of reasons at all. Only this SFW could be said to be really ultimate, but it would thereby also be absolutely arbitrary and, from a rational point of view, completely baseless. The search for ultimate causes leads, as we see, to irrational, or rather, perhaps, a-rational choices, that is, choices made without any reasons whatsoever. And it is certainly hard to say that an agent can be responsible for such choices, no matter whether they are undetermined. Again, ultimacy of source and rational control pull in opposite directions, and ultimate control does not seem to be possible in an indeterministic context.

This is a rather abstract way of presenting the dialectic involved in the incompatibilist demand for ultimate responsibility or ultimate control as a foundation for moral responsibility. But the same feeling of vertigo arises if, in the context of our worry about moral responsibility, we start to ask for the origins of our decisions and actions. The further we move back in our life histories towards our first choices, the less we find ourselves facing them with a definite character and value system, and so with rational and reflective control. And, correspondingly, the more dubious our moral responsibility, in the sense of true desert, becomes with regard to these choices and actions. Some of them, however, are plausibly seen as decisive steps in the forming of our own selves and characters; and thereby the doubt about our moral responsibility spreads over any other choice and action of ours.

Suppose, however, as Kane seems to do, that, in order to avoid this problem about rational control, we hold that ultimate responsibility does not require that the character and motives with which we face SFW situations are actually produced by previous SFWs, and that it is enough that the choice in such situations be influenced, though not causally determined, by our reasons for it. This might help with the problem about rational control, but it would open the door to the problem of ultimacy of source. To see this, consider that, though our reasons do not determine our choice, they causally contribute to it. Plausibly, in a moral SFW, for example, our choice would be different if we had a different set of moral reasons. But then the question about the origin of our actual set of moral reasons becomes important for the question whether our choice has its ultimate source in us. And, as happens with compatibilist views of freedom, it seems that this construal of SFWs is compatible with the agents' reasons not having their ultimate origin in the agent herself. If so, even if the agent could be said to have rational control over her choice, she could not be said to be its ultimate source or have ultimate responsibility for it. And if these ultimacy conditions are required for moral responsibility, as incompatibilists hold, moral responsibility will be undermined. In fact, Ishtiyaque Haji and Stefaan Cuypers have recently argued that several versions of libertarianism, including Kane's, are also exposed to the problem of CNC manipulation (cf. Haji and Cuypers 2001). Even in indeterminist situations, if the agent's values and reasons have their origin in something like CNC manipulation, our intuitions would speak for her lack of moral responsibility, as they did when we raised the problem of Brave New World cases in analysing compatibilism. Nothing less than radical self-formation would seem to suffice to avoid the possibility of CNC manipulation and to ensure that the agent is the ultimate source of her choices and actions; but radical self-formation should start with absolutely baseless and blind choices, for which an agent could not fairly be held responsible, nor be said truly to deserve praise or blame. This would erode the very foundation of her moral responsibility for any of her subsequent choices and actions. Once again, ultimacy and control seem to conflict with one another.

This problem becomes more pressing for Kane given his Humean, or internalist, view of motivation, according to which a motivating reason has to include a desire. We have seen that, for him, even if we rationally see that something is our duty, we shall not do it unless we also desire to do our duty (cf. Kane 1996:207). Desires, however, are paradigmatic cases of states that cannot be directly controlled by our will. I cannot have a desire just by deciding to have it. Think of the desire to do our moral duty. In what sense could we have ultimate control over this desire? It does not seem to be able to arise in us

due to SFWs in which we choose on moral reasons, for, in terms of Kane's Humean, internalist view of motivation, we would not choose on moral reasons unless these included that desire itself to begin with. We simply happen either to have that desire or not to have it. Kane considers reasons to be, essentially, reasons for acting or choosing. These reasons, he insists, cannot motivate unless they include a desire. But, as Parfit points out, "the most important reasons are not merely, or mainly, reasons for acting. They are also reasons for having the desires on which we act… On desire-based theories, any chain of reasons must end with some desire that we have no reason to have" (Parfit 1997:127–8). This is precisely the case with Kane's view of reasons. So it seems that, in order to develop virtuous habits and dispositions, as well as a virtuous character, through our moral (and prudential) SFWs and SFAs, we are, in the end, at the mercy of luck, that is, we depend on whether we happen to have desires to act on moral (or prudential) reasons. We cannot have rational or voluntary control over these springs of our choices and actions. On this Humean view, the roots of our character and reasons are beyond our possibilities of rational and volitional control, let alone ultimate control. So, for someone who wants to show that ultimate responsibility is possible, adopting a Humean view of motivation would not seem to be good advice.

This array of problems surrounding Kane's SFWs also affects the way in which he could deal with cases such as Van Inwagen's judge. This is plausibly a token of what, in Chapter 2, we called "Luther cases". The general worry raised by this sort of case concerns the libertarian categorical interpretation of the alternative possibilities requirement for moral responsibility. In the example of the judge, we see that his deciding to raise his hand, thus granting clemency to the criminal, would seem to be arbitrary and irrational *given the same process of deliberation that led him not to raise his hand*. Now, Kane can accept this judgement. The judge's is a case of what he calls "one-way" rationality, voluntariness and voluntary control, in which the agent's will is settled in one particular direction before the choice is made. The judge's choice has a sufficient explanation in his character, values and motives. In this sort of case, Kane's picture of the agent's psychology need not be pre-Frankfurtian and pre-Watsonian. He can allow the agent to face the choice situation armed with second-order volitions and a system of values according to which she can be said to want to act more on the reasons in favour of one alternative than on the reasons in favour of the other. Kane can accept that the agent, in a situation of this sort, may be ultimately morally responsible for her choice and action. She will be *provided that she is personally responsible for the character, values and motives which explain that choice*. This means that these factors must causally trace back to previous, regress-stopping SFWs of the agent, for which she is directly responsible. But then, the success of Kane's response to the objection posed by "one-way" rational and voluntary choices depends on whether an agent can be said to have ultimate "two-way" (plural) rational and voluntary control over her SFWs. The problem, so to speak, shifts one step back. And, given what we have been arguing about SFWs, it is dubious, to say the least, that agents can actually have such control over them.

Let us now look at Mele's version of the "Mind" objection, applied to an agent's SFWs. Suppose that an agent, S, faces a moral SFW situation where she is torn between two incommensurable sets of reasons, moral and non-moral, and that she finally chooses and acts on the moral reasons. Then imagine a close possible world in which a twin of S's, S*, exists and faces exactly the same situation. Since, according to libertarianism, the

choice is undetermined, S* can choose and act, instead, on the non-moral set of reasons. Suppose that she does. Since S and S* are psychologically identical until the moment of choice, there is no explanation of the difference between their respective choices, which leads to the conclusion that this difference is actually a matter of luck or chance. So S's actual choosing and acting on moral reasons would also seem to be a matter of luck or chance and, if so, she cannot rightly be held morally responsible (praiseworthy, in this case) for it.

Kane's answer to this version of the "Mind" objection is based on his contention that, in SFWs, the agent's effort of will is indeterminate. He writes:

> With indeterminate efforts, exact sameness is not defined… So one cannot say of two agents that they had exactly the same pasts and made exactly the same efforts and one got lucky while the other did not. Nor can one imagine the same agent in two possible worlds with exactly the same pasts making exactly the same effort and getting lucky in one world and not the other. Exact sameness (or difference) of possible worlds is not defined if the possible worlds contain indeterminate efforts…
>
> (Kane 1996:171–2)

If exact sameness of possible worlds or life histories is not defined, the "twin" version of the objection cannot get off the ground. It is not fully clear to us what the contention that efforts of will are indeterminate amounts to. Kane himself acknowledges that "indeterminate efforts are mysterious" (Kane 1996:151), though he thinks that they are no more mysterious than quantum indeterminacy. When Kane introduces this contention in his book, he illustrates it with the analogy of a subatomic particle, whose position and momentum are not both determinate while it moves towards a thin atomic barrier. But Kane himself acknowledges that this is only an analogy, and that our will efforts can be more plausibly seen to correspond to neural macro processes in our brains (cf. Kane 1996:128). He tries then to explicate the efforts' indeterminacy by viewing neural processes from the perspective of a combination of quantum theory, chaos theory and non-equilibrium thermodynamics, as we have seen in the preceding section. In this picture, it is the underlying quantum nature of neural processes that introduces indeterminacy into the brain functioning. Chaos effects amplify this indeterminacy, taking it to the neural level, while non-equilibrium thermodynamic conditions make neural processes highly sensitive to small changes occurring in their environments. Some complex neural processes, in which two neural paths, corresponding to two different and competing sets of reasons, tend simultaneously to reach a certain activation threshold are, experientially considered, efforts of will. The whole nature of the process makes it undetermined which of the two paths will reach the surface, and so which choice the agent will make.

From a metaphysical point of view, this picture would seem to support Leibniz's thesis that there cannot be two things that are exactly alike (indiscernible). This includes worlds, individuals, life histories and efforts of will. But this thesis, unlike the converse thesis that two things that are not indiscernible cannot be one and the same (identical), has always raised suspicions. It is far from being self-evidently true. Suppose, however, for the sake of argument, that Kane is right about the nature of neural processes and that, as a consequence, there are no two numerically distinct and indiscernible such processes.

If so, there are no two indiscernible worlds, individuals, life histories or efforts of will. This, however, would seem to apply to two actually existing items. But possible worlds, individuals or life histories are stipulated, not discovered. It is not clear, then, that Kane's thesis holds across possible worlds. It would not seem to be any contradiction in imagining a possible world that is exactly like the actual world, including the quantum processes that take place in the latter. Logical possibility would seem to be all that is needed to reject Kane's response to the "twin" version of the chance or luck objection to incompatibilism. We do not need physical possibility as well.

However, even if Kane's response were successful against the "twin" version, it does not seem that it could also dispose of the "Rolling-Back" version of the objection. This version does not feature two (real or possible) numerically distinct individuals, but only one. Applied to SFWs, it stipulates that, after this individual makes a certain decision, the world is (repeatedly) reverted by God to the state it was in, say, one minute before that decision and then it is allowed by Him to proceed "forward" again. Since, by the incompatibilist's lights, the decision is undetermined, there will be differences between the original case and some of the replays, as well as among replays, while the process that leads to the decision is exactly the same in the original case and in all the replays. The conclusion, again, is that, in the original, actual case, the agent's decision is a matter of luck or chance.

Suppose, however, that Kane insists that, given the underlying quantum processes in the actual world, the notion of exact sameness is still not defined in this case, so that there is no saying that, in any of these replays, the process leading to the agent's decision will be exactly the same as in the original case, or the same as in any other replay. However, even if Kane were right, there would seem to be a way of construing the "Rolling-Back" objection that is immune to this response. Imagine, then, that, instead of God's making the world return to the state it was in one minute before the decision is made, He makes the world revert to the precise instant that precedes the decision and then allows it to proceed forwards. On this construal, there are not even two numerically distinct processes, real or possible, preceding the decision, which could be compared for sameness or difference, but only one and the same process. Appealing to the idea that exact sameness or difference are not defined for worlds that contain indeterminate events does not cut any ice in this construal. But then, if the agent's decision still varies between the original case and some of the replays, as well as across replays, as it should if it is undetermined, how do we avoid the conclusion that the decision in the actual case really is a matter of luck or chance? As we suggested in analysing the notion of SFWs, Kane's libertarianism would amount to sheer decisionism or voluntarism. Even if it can allow for ultimacy of source, it does not provide enough room for rational control.

Let us finally deal with Kane's speculations about the nature of neural processes. He is not explicit about which sort of position he is adopting for what concerns the mind—body relationship and the related issue of mental causation. However, his texts suggest a view akin to a double-aspect identity theory. He identifies, for example, the indeterministic chaotic process, experientially considered, with the agent's effort of will, and the outcome of that process with the agent's choice (cf. Kane 1996:148). So, in an SFW situation, the physical quantum-cum-chaotic processes that compete to reach an activation threshold are consciously experienced by the agent as her effort of will. An effort of will is, then, that very same physical complex process as it appears to the

agent's consciousness. And a choice, in an SFW situation, is the physical outcome of that process as it is also consciously experienced. In both cases, the indeterministic physical process and its outcome are phenomenologically experienced as mental activity: as a struggle of the agent's will to reach a decision and as a decision or choice, respectively. Physically considered, however, they are mere quantum-cum-chaotic happenings that take place in the agent's brain. But how could one and the same process or event be at the same time an active performance *and* something that merely happens?

There would seem to be two options here. From a broadly naturalistic perspective, phenomenology should give way to the austere view of the physical sciences. The result would be as follows: what is represented in the agent's consciousness as active performances of her own will is, in fact, an array of physical processes and events, subject to probabilistic laws, that take place in her brain. Efforts of will and choices are actually mere happenings illusorily experienced as activities. Or, in some cases, not even so much actually experienced as believed to be experienced. The following text of Dennett's points to the worrying idea that decisions might in fact be happenings rather than acts:

> Once we recognize that our conscious access to our own decisions is problematic, we may go on to note how many of the important turning points in our lives were unaccompanied, so far as retrospective memory of conscious experiences goes, by *conscious* decisions. "I have decided to take the job", one says. And very clearly one takes oneself to be reporting on something one has done recently, but reminiscence shows only that yesterday one was undecided, and today one is no longer undecided; at some moment in the interval the decision *must have happened,* without fanfare. Where did it happen? At Central Headquarters, of course.
>
> (Dennett 1984:80)

On this picture, voluntary and rational control over one's deliberation and decision would also be a phenomenological illusion.

Alternatively, one could accept that the effort of will and the choice are actually as they are (at least sometimes) phenomenologically experienced: they are active performances. And the quantum-cum-chaotic processes in the brain are, as the sciences tell us, mere happenings. But then one should give an account of how these two sets of processes and events relate to each other. They are not identical, since the former are acts or activities and the latter are not. So from this perspective, which allows phenomenology to reveal the real, active nature of efforts of will and decisions, identity claims should be dropped. Suppose, then, that one accepts some form of metaphysical dependence or supervenience of the mental acts on the physical happenings. In this case, there seems to be no way in which the mental acts can influence the physical happenings. Rather, which decision the agent makes will depend, or supervene, on which of the indeterministic processes going on in her brain does actually reach the activation threshold in virtue of quantum and chaotic changes beyond anyone's control. Suppose instead that someone rejects both identity and supervenience. Then she will plausibly have to embrace a form of mind—body dualism, and then she is committed to explaining how the mental acts can exert some sort of causal influence over the physical processes related to physical behaviour.

Kane would reject this latter option. Accepting it would commit him, against his own programme, to a view of free will quite at odds with "the scientific picture of the world", which he thinks any reasonable libertarianism in the modern age should try to respect. But, in any form of materialism about the mind, the quantum indeterminacy of the underlying physical processes will put the agent's choice beyond her voluntary and rational control, as the advocates of the "Mind" objection against incompatibilism have always sustained. Quantum changes suffered by a subatomic particle are not under anybody's control. But then, whether efforts of will and choices are identical to, or supervene on, those quantum processes, they are not under anybody's control either. In fact, if our deliberations and decisions are identical to, or supervene on, quantum-cum-chaotic processes, then we are not better off, for what regards our control over our choices, than if we had in our brains what Pereboom has called a "randomizing manipulator who spins a dial that will land on one of two positions" (Pereboom 2001:53).

So an examination of Kane's hypotheses about the nature of neural processes takes us to the same conclusion as we draw from an analysis of his central philosophical concepts, namely that indeterminism rules out the possibility of an agent's control, and a fortiori ultimate control, over her choices and actions.[1]

We have not examined other versions of libertarianism, especially agent-causation libertarian approaches.[2] Unlike Kane's or Van Inwagen's libertarian theories, which work with a standard construal of causal relations as holding between events (or perhaps states of affairs), agent-causation theories postulate, in addition to these standard causal relations among events, an irreducible causal relation between an agent and certain events, for example certain events in her brain. At the origin of any process ending in a free action there is the agent herself who initiates the process. This move may give some sense to the idea that the agent is the ultimate source of her free actions, but it is hard to accept that it can also deal with the requirement of rational and volitional control over one's choices. Agent-causation libertarian views do not seem to be better off than Kane's theory on this account. Leaving aside the traditional charges of obscurity or mystery against these theories, we cannot see how they can avoid the chance objection traditionally directed against libertarianism in general. The "twin" and the "Rolling-Back" versions of this objection would seem to be as damaging to these theories as they are to Kane's approach. Van Inwagen (cf. Van Inwagen 2000:16–17) has convincingly argued that the example of Alice, who faces a choice between lying and telling the truth, leads to the same conclusion as in event-causal libertarian accounts if we construe that example from an agent-causation perspective, that is, in terms of Alice's causing some brain events which result in bodily movements that constitute her telling the truth (or lying). Supposing again that God produces several replays of the initial sequence (in which, let us assume, Alice tells the truth), and that this again shows the ratio between the two possible outcomes, lying and telling the truth, to converge on some intermediate value, the conclusion that what she actually does is a matter of chance seems equally unavoidable: "If it is undetermined whether Alice will tell the truth or lie, then—*whether or not* Alice's acts are the result of agent-causation—it is a mere matter of chance whether she will tell the truth or lie" (Van Inwagen 2000:17). But then, in our terms, how

could Alice be said to have rational and volitional control, and, a fortiori, ultimate rational and volitional control, over her choice? Agent-causes do not fare better on the issue than Kane's agents. Ultimate control, then, seems clearly incompatible with indeterminism.

Conclusion

In this chapter we have argued that incompatibilism cannot make sense of ultimate control; that, on the assumption of indeterminism, ultimate control is not possible. In the preceding chapter we argued in favour of the incompatibilist contention that ultimate control is required for moral responsibility, understood as true desert. As a result, we now have a strong case for the truth of SMR's premise C, namely that, if determinism is not true, moral responsibility is not possible. And, as a result of the preceding chapters, we also have a strong case for the truth of SMR's premise B, namely that, if determinism is true, moral responsibility is not possible either. The reader will remember that this contention was supported by two lines of argument. One of them rested on two premises: the necessity of alternative possibilities for moral responsibility and the incompatibility of determinism with alternative possibilities. The other rested on two premises as well: the necessity of ultimate control for moral responsibility and the incompatibility of determinism with ultimate control. Granted the truth of premise A, we now have a strong case for the sceptical conclusion of SMR, namely that moral responsibility is just not possible.

5
Overcoming scepticism?
Belief and moral responsibility

As a result of the preceding chapters, we now have a very strong case for scepticism about moral responsibility. In this last chapter, we shall try to see whether this scepticism could ultimately be avoided. Even if it could, however, the strength of the case for it should already lead us to suspect that the scope of moral responsibility, in the sense of true desert, might be narrower than we acritically tend to assume, and make us more cautious about too indiscriminate and profuse applications of this concept. Even if some persons, sometimes, truly deserve blame for their actions, it may well be that the cases in which they do are actually fewer, and their blameworthiness less, than we naturally tend to think. Those who seriously follow the paths of scepticism are likely to develop a more tolerant stance towards the weaknesses of human beings and a reluctance to emit rash judgements about them. If we go deeply enough into scepticism, this will not leave us in the state of mind we were in at the beginning of our journey, even if we are finally able to overcome the sceptical predicament. I think this is a healthy consequence that can improve our lives and bring serenity and peace into them.

It is good that going through the sceptical paths can have such healthy consequences; for it may well be that there is no decisive refutation of scepticism about moral responsibility. I certainly cannot commit myself to offering such a refutation. But I can offer some rather detailed reasons for thinking that scepticism may not be the last word about moral responsibility. Let us draw the general lines of our proposal, starting with an attempt to diagnose the roots of this scepticism.

Scepticism: diagnosis and outline of an anti-sceptical proposal

The main argument for scepticism about moral responsibility, SMR, is logically valid. If we want to deny its conclusion, we must deny at least one of its premises. We have seen two lines of argument in favour of SMR's premise B, that is, the incompatibility between determinism and moral responsibility. The first line leads to premise B through the necessity of alternative possibilities for moral responsibility and the incompatibility of determinism and alternative possibilities. The second line leads to premise B through the necessity of ultimate control for moral responsibility and the incompatibility between determinism and ultimate control. The soundness of either line of argument is sufficient to establish SMR's premise B. This premise is, so to speak, logically overdetermined. However, concerning SMR's premise C, namely the incompatibility between indeterminism and moral responsibility, there is no argument that leads to its truth on the basis of the necessity of alternative possibilities, given that they are clearly compatible

with indeterminism. The main line of argument that leads to SMR's premise C goes through the necessity of ultimate control for moral responsibility and the incompatibility between indeterminism and ultimate control. This line of argument, as we see, has one premise, namely the necessity of ultimate control for moral responsibility, in common with the second line of argument for SMR's premise B. Combining these two lines, we get a unified argument for scepticism, as follows: ultimate control is necessary for moral responsibility; ultimate control is incompatible with determinism and also with indeterminism; therefore (on the assumption that either determinism or indeterminism must be true) moral responsibility is not possible. To fill in the details of this argument, remember that, in the preceding chapter, we distinguished two aspects or components of this condition, namely ultimacy of origin or source and rational control. Suppose that determinism holds. Then, whether or not rational control can be made sense of, ultimacy of source cannot be satisfied, for there are no ultimate causes or origins in a deterministic world. Suppose that indeterminism holds. Then we can have events that, being undetermined, can play the role of ultimate causes, but now the problem, as we have seen, is that it seems that these ultimate causes cannot be under the agent's rational control. In either case, it seems that the requirement of ultimate control cannot be met.

This dialectical situation suggests that the requirement of ultimate control plays a central role in supporting scepticism about moral responsibility. If ultimate control is incompatible with both determinism and indeterminism and is required for moral responsibility, the sceptical conclusion follows. This incompatibility suggests that something may be wrong with this condition. And in fact several authors have argued that the condition raises an impossible demand. If this claim is true, no wonder it cannot be satisfied, whether determinism is true or not. An important argument for the impossibility of ultimate control was developed, some years ago, by Galen Strawson. Strawson uses the expressions "true responsibility" or "true self-determination", but the content of these expressions is arguably equivalent to Kane's Ultimate Responsibility and to what we have called "ultimate control". Kane himself writes that "Strawson's 'true responsibility' [is] what I designate as 'ultimate responsibility'" (Kane 1996:34). A consideration of Strawson's argument will be of some help in elaborating a more detailed diagnosis of the roots of scepticism. Strawson's argument runs as follows:

(1) Interested in free action, we are particularly, even if not exclusively, interested in rational actions (that is, actions performed for reasons), and wish to show that such actions are or can be free actions. (2) How one acts when one acts rationally (that is, for a reason) is, necessarily, a function of, or determined by, how one is, mentally speaking… (3) If, therefore, one is to be truly responsible for how one acts, one must be truly responsible for how one is, mentally speaking—in certain respects, at least. (4) But to be truly responsible for how one is, mentally speaking, in certain respects, one must have chosen [consciously and explicitly] to be the way one is, mentally speaking, in certain respects… (5) But one cannot really be said to choose, in a conscious, reasoned fashion, to be the way one is, mentally speaking, in any respect at all, unless one already exists, mentally speaking, already equipped with some principles of choice, *"P_1"*—with preferences, values, pro-attitudes, ideals, whatever—in the light of which one chooses how to be. (6) But then to be truly responsible on account of having chosen to be the way one is, mentally speaking, in certain respects, one must be truly responsible for one's

having *these* principles of choice, P_1. (7) But for this to be so one must have chosen them, in a reasoned, conscious fashion. (8) But for this, that is, for (7), to be so one must already have had some principles of choice, P_2, in the light of which one chose P_1. (9) And so on.

(Strawson 1986:28–9)

Strawson's conclusion is that "true self-determination is logically impossible because it requires the actual completion of an infinite regress of choices of principles of choice" (Strawson 1986:29). The ultimacy requirement that we found in Kane is clearly echoed in premises (2) and (3) of Strawson's argument: a subject's rational actions are explained by her mental constitution, or character; so, if a subject is to be truly (ultimately) responsible for the way she acts, she has to be truly (ultimately) responsible for this mental constitution, or character, by having *chosen* it. The rational control requirement appears in this argument, from premise (4) onwards, as a condition of true (ultimate) responsibility: to be truly responsible for one's mental constitution one has to have chosen that mental constitution in a *rational* way, that is, in the light of principles of choice. The argument is highly instructive. It shows why those two aspects of ultimate control, namely ultimacy of source and rational control, might not be jointly satisfied. We can make a rational choice of a certain action on the basis of our reasons, but the choice cannot at the same time be an ultimate source of the action, because it is a function of what reasons we have and find convincing and this, in turn, depends on our character, on how we are, "mentally speaking". But the same applies to a rational choice of our character. **On** the other hand, we can choose a way of acting, or a way of being, in such a way that the choice is an ultimate, underived source of such an action or character, but the choice is bound to be made on no basis at all and so in a completely arbitrary way: it cannot be a rational choice. It seems, then, that no choice can be both an ultimate and a rational source of one's actions. In the end, we are bound to act on the basis of factors that we cannot have rationally chosen and for which we cannot be truly responsible. Another way of putting Strawson's point about the impossibility of actually completing an infinite series of choices of principles of choice is to say that ultimate control involves a self-defeating demand for self-creation. As Randolph Clarke has put the point, according to Strawson "rational free action would be possible only for an agent who was *causa sui*" (Clarke 1997:37; cf. also Levy 2001). Ultimate control (true responsibility, ultimate responsibility, true self-determination) is a requirement that cannot be met.

But if ultimate control is so plainly impossible to satisfy, is it not unreasonable to raise it to the status of a requirement for moral responsibility? Strawson voices a related objection to his contention on behalf of an opponent to it: "It may be objected that the kind of freedom this argument shows to be impossible is so obviously impossible that it is not even worth considering" (Strawson 1986:30). And his reply is that "the kind of freedom that it is an argument against is just the kind of freedom that most people ordinarily and unreflectively suppose themselves to possess" (Strawson 1986:30). We shall come back to this reply. Whether or not it is finally true, there is certainly something to it. We can develop this point in a slightly different way. The requirement of ultimate control arises out of the very nature and scope of moral responsibility attributions. From an objective, third-person point of view, a serious attribution of moral responsibility is directed to the agent herself, on the assumption that she is the true origin of the action, or

consequence thereof, for which she is held responsible. These attributions have a deep and strong effect on our worth and value, as well as on our self-esteem and sense of dignity. It is understandable, then, that we want to keep personal control over the grounds on which such attributions are made. Otherwise, we might find our worth increased or diminished without our participation, in quite unexpected and hazardous ways. This is why we want to have, and think we actually have, ultimate control over those things for which we are rightly held responsible. In fact, as we have already pointed out, this desire for, and belief in, control over the responsibility we bear for what we do is also at the root of the alternative possibilities condition. If we are to be morally responsible for a certain action, we understandably want to have freedom to do that action or do something else instead, including simply not doing that action. If this is not the case, it seems that we lose control over our moral responsibility, for then we might be morally responsible for actions that we could not avoid performing. This common root of the ultimate control and the alternative possibilities requirements for moral responsibility leads naturally to unifying them, as Kane does in his notion of ultimate responsibility. A useful label for this unified requirement, inspired by Fischer's terminology, which we have already used, is "ultimate regulative control".

These reflections speak up for the existence of a constitutive link between moral responsibility and some deep, ultimate form of control. We have argued in favour of such a link against compatibilist denials of it. Without ultimate control, it seems, we can still give some sense to the expression "moral responsibility", but not the crucial sense of true desert, of true praise- and blameworthiness which, being a central aspect of our intuitive, pre-theoretical idea of moral responsibility, is also (and thereby) a central concern of philosophical reflection about it. Without the assumption of some form of deep, ultimate control over our actions and choices, attributions of moral responsibility, of moral praise- or blameworthiness, would come close to aesthetic judgements of people on the basis of their innate physical characteristics, such as their beauty or ugliness, or of some of their psychological capacities, such as their intelligence or their talent for music.

However, to acknowledge the necessity of a deep, ultimate form of control as central to our intuitions about moral responsibility, understood as true desert, still leaves open what the right content of that condition is. Accepting the intuition that some form of deep control over one's actions is a requirement for moral responsibility does not imply accepting a particular theoretical construal of that requirement. It is not obvious, then, that Strawson is right when he holds that his "true responsibility" coincides with "the kind of freedom that most people ordinarily and unreflectively suppose themselves to possess", or, in other words, with our intuitions about deep, ultimate control as a condition of moral responsibility. It may well be that deep, ultimate control looks impossible in virtue of a particular theoretical construal of such a condition which, on closer inspection, might be shown not to reflect our intuitions about the requirements for moral responsibility. It may well be that a philosophical construal of the idea of ultimate control is available which respects our intuitions about the necessity of that condition for moral responsibility while, at the same time, avoiding the traits of current philosophical proposals that make ultimate control appear to be an impossible demand. We shall argue that such a construal may actually be available and indicate the direction in which it might be framed. This will constitute the positive, curative part of our proposal. Before that, however, we still need to get a more precise diagnosis of the roots of scepticism

about moral responsibility. We still need to investigate whether our everyday intuitions about the conditions of moral responsibility, "the kind of freedom that most people ordinarily and unreflectively suppose themselves to possess", actually raise an impossible demand or whether this impossibility is rather the artefact of particular theoretical accounts of those intuitions. Our conjecture is that the latter is true, but not the former, and that what makes a deep, ultimate form of control appear impossible in the context of these theoretical accounts is *their almost exclusive emphasis on the will and will-related acts, particularly choices*. Our proposal, to anticipate, will be that ultimate control should be construed in terms of beliefs rather than choices, and that, with such a construal, the condition might actually be met.

This almost exclusive emphasis on the will and will-related conative acts, especially choices, as the core of ultimate control can be clearly seen in current philosophical approaches to this condition, such as Robert Kane's or Galen Strawson's. Think, for example, of the crucial role that self-forming willings play in Kane's notion of ultimate responsibility. For him, an agent can only be said to have ultimate responsibility (control) over her actions by virtue of her performing causally undetermined, plural rational and voluntary choices (self-forming willings) through which she originally builds up and brings about her own character and motives, the springs of those actions. So the very root of ultimate control over one's actions is an act of will, a self-forming willing or choice. In Strawson's case, the crucial role of choice is also undeniable. True responsibility for one's actions requires true responsibility for the mental constitution they arise from, and this, according to Strawson, implies that one has *chosen* that mental constitution as well as the principles on which such a choice is made. It seems, then, to be an unexamined assumption of Kane's and Strawson's approaches that there is a constitutive link between ultimate control and the will: it is this will-centred view of ultimate control that, in the end, justifies considering an agent as the genuine, ultimate author of her actions, thereby grounding moral responsibility attributions.

Now we shall contend that this assumption of a constitutive link between deep, ultimate control and the will, in the form of willings or choices, unlike the assumption of a constitutive link between responsibility and deep control, is not backed by our intuitions about when an agent truly deserves praise and blame for some of her deeds. I think that we are prepared to acknowledge that an agent is the ultimate source or origin, the true author and creator of a certain performance of hers, so that she fully, unrestrictedly deserves praise or blame for it, even if we do not see that performance, in any important sense, as a result of the agent's choices or acts of will. If this is true, not all control, even ultimate, is dependent on choices or acts of will. And we strongly suspect that the opposite assumption might be at the root of scepticism about the possibility of moral responsibility. Suppose, in fact, that any item over which an agent has ultimate control has to depend, in the end, on a choice of hers. And add to this the completely plausible assumption that, for it to ground the agent's responsibility for that item, the choice has to be rational: it has to be made on the basis of reasons or principles that justify and explain it. If so, according to the first assumption, she has to have chosen those reasons or principles, which in turn, on the second assumption, implies that she must already have another set of reasons or principles for the choice…and we have started the infinite regress of choices of principles of choice that Strawson rightly claims to be impossible to complete. However, if we drop the first assumption, the regress does not need to start.

In fact, we shall argue that, in many cases in which we acknowledge an agent's full authorship, and so full praise- and blameworthiness, for an accomplishment of hers, it is precisely the fact that she has *not* chosen either the principles on which such an accomplishment depends or many other factors that have made it possible that makes us see the agent as truly deserving our praise (or blame) for it.

Much of the reflection on free will and moral responsibility has been based on two related, and questionable, assumptions. We have already indicated the first: it is that the will, in the form of conative acts, is the last foundation of moral responsibility. We have found this assumption in Kane and Strawson, but its tracks can be followed in many other authors. An enormous weight has been placed on the instant of decision or choice, with a corresponding oblivion of the cognitive context in which choices are made. The picture of responsible agents as constantly making decisions is quite likely distorted. I agree with Linda Zagzebski when she writes that "choice even in the realm of action is much less important and occurs much less often than is commonly believed" (Zagzebski 2000b:212). Most of the time we act on the basis of our beliefs, of our cognitive view of things, which includes our evaluative beliefs about what is valuable and worth pursuing, with no conscious decision. And in cases in which decision is involved, when the instant of decision comes, much of the matter is already settled, on the basis of such cognitive views. In contrast, it is instructive to consider the view of choice that derives from Kane's conception of self-forming willings in the context of a response to Frankfurt cases (cf. Kane 1996:192). According to Kane, in a Frankfurt case, a counterfactual intervener (Black, say) would not be able to control the agent's choice provided that this is a self-forming action or willing. The reason is that, no matter how the process prior to the choice goes, Black will not be able to predict which final decision will be made. Watching the agent's deliberation process and collecting a good deal of information about it, Black may be confident that the agent will finally decide in the way he wants her to and so allow the agent to make the choice fully herself. The agent's choice, however, may contravene Black's expectations. Alternatively, Black can wait and see what the agent actually chooses, but then it will be "too late to control the choice" (Kane 1996:192). This picture of choice corresponds, we think, to the assumption we are referring to. Choice, as the ultimate foundation of moral responsibility, appears to arise as something unpredictable, as an event that suddenly breaks the process of deliberation, leaving a would-be controller with no crack in which to insert his wedge. It is clear how this picture can easily fall prey to versions of the "Mind" argument. As we have suggested, it is this view of moral responsibility as ultimately dependent on choice that makes ultimate responsibility or control appear to be an impossible demand. In our response to Frankfurt's attack on PAP, we were careful to avoid such a picture, insisting instead, against Pereboom and Zagzebski, on the importance of cognitive possibilities, such as thinking of certain moral, evaluative reasons, and on the constant availability of alternative actions even in the presence of a Frankfurtian, counterfactual controller.

The second assumption we referred to above follows naturally from the preceding one. It can be seen most clearly in Strawson's claim that, in order for true responsibility to be possible, an agent has to have chosen the principles on which she makes her choices. Unless she has chosen those principles, she cannot be truly praise- or blameworthy for any of her actions. This claim reveals a deeply individualistic view of human agents as radically self-made, self-contained entities, whose constitution does not owe anything to

factors external to them, or at least it cannot do so except at the cost of their not truly deserving praise and blame. Nothing short of absolute, radical origination, which includes the choice of one's own reasons, can be enough for ultimate control, and so for moral responsibility. So, if the explanation of a person's acts traces back to factors external to her self and so to her possibilities of choice, her moral responsibility for those acts is jeopardized. On this deeply individualistic view of human agents, the fact that they appear to be socially constituted is seen as threatening to their moral responsibility. The compatibilist reaction to this is to claim that agents can be morally responsible even if their actions' origins go far beyond the agents themselves. The incompatibilist reaction is that they cannot. Libertarians, then, try to show how it is possible for an agent to be a radical, absolute source of her actions in spite of social influences on her character and motives. Non-libertarian incompatibilists hold that, given that human beings are socially constituted, the radical origination that would be required for moral responsibility is not possible at all. But the right reaction, which we shall be arguing for, is that an agent can have ultimate control of her actions and be her deep, ultimate source partly by virtue of being socially constituted. In other words, the social nature of human beings is best viewed as an enabling factor rather than an obstacle for her moral praise- and blameworthiness.

In the light of all these considerations, we can now outline the essentials of our positive proposal. Our general position can be boldly stated as follows. Ultimate regulative control is the freedom-relevant necessary and sufficient condition of moral responsibility, understood as true desert. We have defended the alternative possibilities condition (the "regulative" aspect of control) against several criticisms of it. We have also argued for the ultimacy aspect of control against compatibilist attempts to reject it. But we have not rejected important compatibilist insights about (rational) control. In its attempt to construe the notion of moral responsibility without resorting either to alternative possibilities or to ultimacy, compatibilist reflection on moral responsibility has brought to light other aspects of that notion which might otherwise have gone unnoticed. Following Kane, we have accepted the necessity of some of those aspects for moral responsibility. However, we have insisted that regulative control (alternative possibilities) and ultimate control are also required in order to have a sufficient condition. This contention puts our proposal on the incompatibilist side. If some form of the Consequence Argument is sound, alternative possibilities are not compatible with determinism, though they clearly seem to be compatible with indeterminism. The hardest problem, of course, comes from ultimate control. Our purpose is to show that ultimate control is indeed possible and that, even if it is not compatible with determinism, it may nevertheless be compatible with indeterminism. We shall therefore try to reject the incompatibility between ultimate control and indeterminism. This incompatibility, as the reader will recall, is an essential step in the main argument that supports SMR's premise C, namely that, if determinism is not true, moral responsibility is not possible. On this basis, we shall try to reject SMR's premise C so as to avoid the sceptical conclusion of this argument.

In order to accomplish this anti-sceptical program we shall be contending that, though the sort of control required for moral responsibility may be voluntary control, or control based on decisions or choices, it need not be. Not all control is voluntary control. The opposite assumption has led some writers to sever the link between responsibility and

control, which we think is constitutive and non-negotiable. So, for example, in a recent book (Owens 2000), David Owens has claimed, on the basis that we do not have voluntary control over our beliefs, that the notion of responsibility, as well as related deontological notions, applies only to actions, and not to beliefs. We shall argue that this view is not correct and that, though we agree that belief is not voluntary, we can rightly be praised or blamed for our beliefs, for we have over them a form of control that, though not essentially based on the will, is none the less fit to support praise- and blameworthiness attributions.[1]

As we anticipated, instead of a will-centred account of moral responsibility we recommend a belief-centred perspective. At the foundational root of moral responsibility we shall not place conative phenomena, such as choices, decisions or (first- or second-order) volitions, but cognitive ones. Especially important will be a distinctive class of beliefs, namely evaluative beliefs about what is really valuable and worth pursuing or avoiding in life. Beliefs of this sort play a central role in explaining those of our decisions and choices that have moral import. We recommend, then, a cognitive rather than a conative conception of moral responsibility.

This cognitive turn, as it might be called, in the theory of free will and moral responsibility, is not without precedents. We already know some of them. A step in this direction is Gary Watson's emphasis on values, versus Frankfurt's second-order volitions. A more detailed contribution is Susan Wolf's Reason View, according to which the sort of freedom that is relevant to moral responsibility is the ability to act on one's values *and to form those values* in the light of an appreciation of the True and the Good. Against Wolf, however, we shall contend that ultimate control, which she finds intuitively appealing as a requirement for moral responsibility though in the end impossible, is instead possible, and also compatible with a cognitive approach, based on evaluative beliefs, to moral responsibility. So, unlike Wolf (and Watson), we shall develop our account in a libertarian direction rather than in a compatibilist one. Also significant, though not dealt with in the previous chapters, are Paul Benson's and Henry Richardson's contributions (Benson 1994, Richardson 2001), in which they insist on a substantive approach to freedom and autonomy, with an emphasis on the *content* of the agents' attitudes, against purely formal or structural approaches.[2] In fact, we can also find this emphasis on content in Wolf's conception, in which she reacts against Watson's formal view of values and Frankfurt's structural perspective.

If we refuse to place conative acts, such as decisions or choices, at the roots of moral responsibility, we also deny the view that moral responsibility ultimately rests upon conative states, such as desires. We reject a Humean view of motivation, which we have seen at work in Kane's libertarian theory, with damaging effects on his project. In this context, we find some of Derek Parfit's anti-Humean remarks in one of his papers quite suggestive and congenial. He speaks, for example, of "truths about what is worth achieving, or preventing" (Parfit 1997:129), which are quite close to our evaluative beliefs. He also writes that "the most important reasons are not merely, or mainly, reasons for acting. They are also reasons for having the desires on which we act. These are reasons to want some thing, for its own sake, which are provided by facts about this thing. Such reasons we can call *value-based*" (Parfit 1997:127–8). In fact, if Humean internalism about motivation is true, then deep, ultimate control over our actions does not seem possible, for, even if I could form correct evaluative beliefs, they would remain

motivationally inert unless they connected with a previously existent desire, a mere fact about me for which I am not praise- or blameworthy. It is important, then, that evaluative beliefs are able to motivate and give rise to desires for which they provide reasons.

Another field of research that may prove useful to our proposal is epistemology. Over the last two decades, some epistemologists have proposed looking at ethics in order to throw light on normative epistemological notions, such as justification. Roughly, a belief is justified just in case it has been formed as it *ought* to have been, where the "ought" is the normative "ought" that we also apply to actions. If this points to the right direction, then there is room for talk about an ethics of belief, and about responsibility, praise- and blameworthiness, in the epistemic and not only in the practical realm. The discussion about this research field is lively and contributions to it have increased enormously. In fact, however, given the strength of the case for scepticism about free will and moral responsibility about actions and the obscurity that surrounds these notions, one might be tempted to think that epistemology is unlikely to get more light about its central notions by looking in this direction. But maybe the right reaction is to encourage interdisciplinary research. Our proposal, in fact, can be seen as somehow the reversal of that general epistemological project, in that we intend to see whether scepticism about moral responsibility for our actions could be overcome, or at least undermined, by focusing on beliefs and the responsibility we bear for them.

Finally, the cognitive approach to moral responsibility that we are suggesting should also avoid the individualistic view of human agents that goes with the will-centred, conative approach. At this point, we also find interesting parallels with some epistemological issues. Cartesianism is driven by the impulse towards an ultimate rational examination and control over anything that is relevant to the process of inquiry, as a requirement of epistemic responsibility (cf. Hookway 1994:214–15). This impulse is quite analogous to that which drives will-centred conceptions of ultimate control as a condition of moral responsibility. According to these conceptions, we cannot have such ultimate control, and so moral responsibility, for our actions unless the springs of those actions are the result of our own choice. It can be seen that both approaches adopt a profoundly individualistic stance towards human beings as cognitive or practical subjects. No epistemically or practically valuable result can be credited to an individual if the explanation of such a result traces back to factors beyond the reach of the individual's powers of rational reflection and choice. An absolute origin in the individual as a self-sufficient, self-contained entity is required of any accomplishment for which she truly deserves praise or blame. The individual may be subject to cognitive and practical influences from her social environment, but she has to make those influences part of herself by means of her own rational reflection and choice if she is to deserve credit for any result that such influences may help to explain. The problem, of course, is that rational reflection and choice require reasons and principles that, in the end, cannot be subject to rational reflection and choice except at the cost of an infinite regress. In both cases, scepticism (about knowledge or about moral responsibility) is the expected result. The question, once again, is whether we can retain the intuitions behind the ultimate control requirement without the costs of will-centred, individualistic construals of that requirement. And a positive look at the social nature of human agents, for what concerns the possibility of their moral responsibility, would seem to be part of such an achievement. We should take seriously the view which Wilhelm Dilthey expressed a long

time ago when he said that a human being is a crossing-point of the large systems of social interaction. Of course, Dilthey is not alone in his insistence on the social constitution of the individual. He writes in what Tyler Burge (Burge 1979) has called the "Hegelian" (as opposed to the Cartesian) tradition, with its emphasis on the role of social objectivities and institutions in the shaping of the individual mind. In this context, it seems right to say that the dominant lines of reflection on free will and moral responsibility have followed the Cartesian, individualistic tradition. Just as, in the philosophy of mind, Burge has favoured a Hegelian perspective on the nature of intentional content against the dominant Cartesian approach, we tentatively recommend such a perspective in the reflection on moral responsibility as well.

Let us now substantiate and develop further the proposal we have just outlined in this section.

Belief, control and responsibility

Let us start by noting that talk about responsibility for one's beliefs not only makes perfect sense, but is quite common in everyday life. Think of such remarks as the following, which implies blame: "How could you believe what he told you? Don't you know how often he lies?" In cases like this, similarly to what happens in the sphere of actions, excuses are likely to be expected: "Well, this time he really looked sincere." Other cases concern formation of belief in the light of clearly insufficient evidence. We sometimes blame people, including ourselves, for being too rash and careless in coming to have certain beliefs. On the positive side, we also praise people for not being too credulous. There are also interesting cases of praising someone for not believing something which has almost overwhelming evidence in its favour. Some thrillers provide good examples of this: think of the police inspector who keeps investigating, against her superiors' direction, when everything seems to point to someone as the culprit of the crime. Depending on many circumstances, we tend to see some of these cases as exemplifying either the virtue of tenacity or the vice of stubbornness.

These are just a few examples of our application of responsibility-related concepts to the cognitive, and not only to the practical, field. How are we to understand this application? Some years ago, Bernard Williams (1973) famously argued that beliefs are not under our direct voluntary control, so that there is not much room for deciding to believe. It is incompatible to have a belief and to know that one has that belief only because one has decided to have it. The main reason for this contention is that belief constitutively aims at truth. As Williams writes: "If I could acquire a belief at will, I could acquire it whether it was true or not; moreover I would know that I could acquire it whether it was true or not. If in full consciousness I could will to acquire a 'belief irrespective of its truth, it is unclear that before the event I can seriously think of it as a belief, i.e. as something purporting to represent reality" (Williams 1973:148).

Williams's rejection of doxastic voluntarism has gathered wide acceptance. I find it convincing too. There are some disagreements, however. In a recent paper (Shah 2002), Nishi Shah contends that, if Williams's argument against doxastic voluntarism were sound, then no activity that has a constitutive aim could be performed at will. Lying, for instance, whose constitutive aim is to deceive, could not be voluntary, which is plainly

false. But this criticism is mistaken. Williams's argument does not rely on the assumption that believing has a constitutive aim, but on the *particular* constitutive aim it has, namely *truth*. Beliefs aim to conform themselves to the way things actually are, whereas, for example, decisions and desires aim to change the world so that it conforms to them. In other words, what puts beliefs beyond the reach of the will is that they have a mind-to-world direction of fit, whereas decisions and desires have a world-to-mind direction of fit. Shah is wrong, then, in holding that "if Williams's argument goes through for the case of belief, then it will go through for any aim-constituted activity" (Shah 2002:438).

In the context of the present research, it is good that doxastic voluntarism is false. For suppose that it is true, so that believing is, after all, an action or activity. Then it is hard to see how appealing to beliefs could be of help in the task of overcoming scepticism about moral responsibility for our actions, for beliefs would then be affected by all the sceptical considerations that we have gone through concerning moral responsibility for actions. Inquiry is a cluster of activities: gathering and assessing evidence, drawing conclusions, etc. And it certainly seems to be true that we praise or blame people for the way they conduct their inquiries, presumably because the way they conduct them increases, or decreases, as the case may be, the probability of having true and justified beliefs. This, however, is a particular case of ascribing responsibility for actions. If responsibility for beliefs derives *entirely* from responsibility for our cognitive activity, we may expect little help from focusing on beliefs for what concerns the possibility of overcoming scepticism about moral responsibility for our actions. If there are sceptical doubts about control and responsibility for actions, the doubts will also extend to those actions related to inquiry and belief formation. So the thesis that the control we may have over our beliefs derives entirely from voluntary control over our cognitive activity is not of much help to our anti-sceptical project. But we might have a different form of control over our beliefs. This control would not be voluntary, but would also not be merely indirect. It would not derive entirely from our voluntary control over our epistemic activity. My suggestion is that we actually have a control of that sort, which grounds some attributions of responsibility for our beliefs.

As a first step towards a characterization of this sort of control, let us think of someone who is carrying out a rather complicated addition without an electronic calculator. She performs the task carefully and gets the right result. She is praiseworthy on both accounts, and has had control over both the process and its result. It is not a matter of luck that she has got it right. But think what having control amounts to in this case. It does not have to do with choices or acts of will in any important sense. The control she has consists rather in her yielding to the internal structure of the thing itself, the figures and the addition rules. It is, so to speak, a passive form of control, which she exercises precisely in being guided by what is there, in the addition problem. She does not choose the rules. In fact, she would *lose* control of the process if she chose the rules (or the figures), and she would rightly be blamed if she did that. She has neither chosen nor created either the rules or the figures. They come "from outside" her self or her will. But this does not exclude her having deep control over her belief about the result of the addition.

This example involves an algorithmic procedure to arrive at the correct belief about the result of an addition. But we can consider more complex cases of belief and belief formation, though still with no specific moral profile. Think of great achievements in the

fields of science or philosophy. They can harmlessly be taken to be systems of beliefs. These cases show, with special clarity, that authorship concerning beliefs may be as important for a person's worth and self-esteem as authorship concerning actions or decisions. Think of the virulence of many discussions, in the past and the present, about the paternity of a certain idea or theory. To mention just a couple of cases, remember the famous dispute between Newton and Leibniz concerning the invention of infinitesimal calculus, or, more recently, the discussion about the true discoverer of the virus causing AIDS. These disputes go deep into questions of personal worth and value. In connection with this, think also of the strongly negative moral assessment that plagiarism raises in most of us. Questions of real source or origin, and of corresponding praise- and blameworthiness, are no less significant and pressing in the cognitive field than in the practical one.

Let us now reflect on a particular example, Newton's *Philosophiae Naturalis Principia Mathematica,* in order to discern some aspects of our notion of authorship or source in relation to cognitive achievements. We agree, I hope, that, provided that he wrote it, Newton truly deserves our unrestricted praise and gratitude for producing such a great scientific work, which has deeply influenced our view of the natural world. We are grateful to him not only for the strenuous effort he expended in producing his work, not only for the careful cognitive activity he carried out in order to produce it, but also for the result itself, for the important ideas and beliefs that such an activity and effort brought about. He has genuine merit and truly deserves praise for his great intellectual achievement. But now consider how many elements that were arguably indispensable for his work to be possible were not of his own making, how much his work is actually indebted to cognitive factors that he did not give origin to. It is hard to see, for example, how the *Principia Mathematica* could have been produced without the contributions of such thinkers as Copernicus, Galilei or Kepler, whose work is in turn indebted to prior scientists and philosophers. And something similar can be said about many methodological and mathematical instruments and empirical data that Newton did not create or discover himself, but found already there. Nevertheless, this does not incline us to question Newton's full authorship, merit and responsibility for his work. But it seems to show that our judgements about authorship, praiseworthiness and responsibility for cognitive achievements do not correspond to the pattern of Kane's or Strawson's conceptions of ultimate control, at least for what concerns the ultimacy of source aspect. If we dig deep into the origins of Newton's work, we shall probably find a rather messy situation, with some aspects ultimately traceable to Newton himself and many others whose roots go far beyond him. But that does not prevent us from seeing his work as a whole, as an articulated system of propositions and laws, as truly and ultimately attributable to him. This suggests that something similar might be the case in the practical realm.

But I think that Newton's example can also make Kane's and Strawson's conceptions of ultimate control wanting for what regards the rational control aspect. Though surely the will, in the form of choices, is involved in the process of creation of the *Principia Mathematica,* Newton's rational control over such a process does not mainly consist in voluntary acts or choices, but, to a large extent, in his passively yielding to the internal requirements and structure of the subject itself, in his sensitivity to the actual relations of deductive and inductive justification, to the internal connections between concepts, to the

perceived necessity of certain steps in the reasoning process. In fact, for him to be legitimately praised for his work, it is centrally important that he produced it with due humility and respect for the data and the principles of valid reasoning, as well as for their relations. If we discovered that he made these aspects depend on his will, our judgement about his merit and praiseworthiness would be drastically affected. This suggests that the link between the control required for true desert and the will, at least in the cognitive realm, is much looser than has traditionally been assumed in the reflection about moral responsibility.

These examples allow some tentative but fairly interesting remarks about certain varieties of control related to true desert.

First, if a particular theoretical construal of the sort of control required for true desert, such as Kane's or Strawson's notion of ultimate control, leads to rejecting the subjects' praiseworthiness for their beliefs in the preceding examples, we should conclude that, at least in the cognitive field, that theoretical construal is not correct, not that our subjects are not praiseworthy, for our intuitive judgement about their praiseworthiness is much firmer than our confidence in such theoretical constructions. They do indeed have all the control over their beliefs that is required for them to truly deserve praise for having them. And if we call all the control required for true desert "ultimate control", they certainly have ultimate control over their beliefs. In the addition example, it would be ludicrous to deny that the agent really deserves praise on the basis that she has not chosen or created the rules of addition, or the figures she has worked with. And the same applies to many aspects of Newton's great work.

Second, although there is some distance from the preceding remarks about true desert and control over *beliefs* to conclusions about true desert and moral responsibility for actions and the possibility of overcoming scepticism concerning them, the examples also offer some clues as to how this distance might be bridged. For simplicity, let us focus on the addition example. The subject's belief about the result of the addition is a state. But her performing the calculation is certainly a sequence of actions. It is cognitive activity aimed at getting the true result of the addition. As it happens, our judgement is that the subject also deserves praise for the way she performs this activity. For the sake of argument, let us call each step in the process of calculation a "choice". In this case, that the subject has rational control over these choices means that she makes them according to the rules of addition. These are, in Strawson's terms, her "principles of choice". Now, if Strawson were right about the control required for true desert, then, in order for our subject to truly deserve praise for her calculation, she should have chosen these principles of choice. But this seems simply wrong. These principles are *the* principles to apply to this activity. So, if our subject is truly praiseworthy for performing her task carefully, according to these principles, this would seem to provide a prima facie counterexample to Strawson's construal of ultimate control as true responsibility or true self-determination, as well as to Kane's construal if, as he himself claims, his "ultimate responsibility" is what Strawson calls "true responsibility". The counterexample is still only prima facie in that the case at hand does not have a specific moral import. So it might still be that *moral* praise- and blameworthiness for actions does require ultimate control in Strawson's or Kane's sense. But the example certainly casts some doubts on the correctness of these conceptions of ultimate control as necessary for moral responsibility.

Third, the examples also offer a prima facie case for thinking that the "external" origin of the factors that explain a certain accomplishment or activity of a subject does not, in itself, provide a reason for holding that the subject is not the true origin of such an accomplishment. In the addition case, both the arithmetic rules and the figures that our subject has to work with have an external origin and are certainly explanatory of both her calculation activity and her belief. As for the rules, she has acquired them through a learning process. She has not chosen or invented them. With due respect for the higher complexity of the case, similar remarks may also be made concerning Newton's example. But if, as we have suggested, we think of an individual as at least partly socially constituted, we do not need to deny that our subjects are the true origin of their activity and their beliefs, so depriving them of the merit they deserve. Departing from a rampant individualism and accepting the social nature of human beings provides a way of reconciling the influence of external factors on a subject's performances and her having ultimate control over them. However, since there are cases in which external factors do actually undermine the subject's ultimate control and responsibility (think, for instance, of Brave New World cases), this leaves open the task of distinguishing between these two possible effects of external factors on the subject's control and responsibility, and of explaining the difference.

Finally, it seems to me that the assumption that our subjects had alternative possibilities (regulative control) underlies our judgement that they are praiseworthy for their activity and the beliefs they arrived at through it. In the addition case, it seems true that the agent could have acted less carefully in adding and arrived at a false or less justified belief about the result of her problem, and that this thought partly explains our disposition to praise her. And the same goes for the example of Newton. But we shall return to this important point in a later section.

The preceding considerations about the nature of the control required for praiseworthiness, which suggest its rather loose connection with the will, can also be seen to apply, *mutatis mutandis,* to the field of literary creation. Even if, in this field, the role of the will and choice may be much larger than in the case of scientific and philosophical creations, we still find control to depend here, to a wide extent, on the author's respect for the internal structure and tendencies of the fictional world she has created. Let us sustain this claim by arguing that one of the features that distinguish great literary narrative or drama from second-rate creations is the author's yielding to the intrinsic, autonomous dynamics of the characters, instead of constantly making decisions on their behalf about what they are to think or do at each moment. Voluntary interference in the life of the characters leads, somewhat paradoxically, to loss of control over the work, so giving rise to the impression of capriciousness and lack of depth that characterizes weak literary creations. On the other hand, and also somehow paradoxically, allowing oneself to be controlled by the intrinsic dynamics of the work is one sign of the author's control over it. Wilhelm Dilthey spoke about the experience of being led in creation, thus referring, as I interpret him, to this yielding to the internal life of the created world that underlies valuable literary works.

The variety of control we are pointing to is not a purely passive stance, however; it is, rather, a form of enlightened humility and respect for the object, which disposes the subject to recognize the proper value of what is other than herself. Far from being at odds with authorship and responsibility, this form of control is a mark of great authors and a

legitimate ground for praiseworthiness in the field of literary creation. With due attention to the differences, something similar can be held to obtain, as we have seen, in the field of scientific and philosophical creation. We shall contend that it is also centrally important in the practical field, in connection with moral responsibility for actions. It is now time to come back to this problem.

Evaluative beliefs and moral responsibility

We have insisted that the intuition behind the idea that ultimate control is a requirement for moral responsibility, understood in the basic sense of true desert, is largely correct. We have expressed doubts, however, about the correctness of prevailing theoretical elaborations of that intuition in terms of acts of will or choices. The crucial problem with the attempt to ground moral responsibility in acts of will or choices is that, in the end, such choices are bound to be made without reasons and so in a baseless, arbitrary way, as Strawson's sceptical argument makes clear. Our proposal is to analyse the notion of ultimate control from a cognitive perspective, in terms of beliefs. However, interested as we are in moral responsibility for our decisions and actions, evaluative beliefs about what is important, worth pursuing and caring about in one's life are especially relevant. Let us develop and justify this proposal further.

Evaluative beliefs are beliefs with an evaluative content. Not any beliefs of this sort, however, are relevant for purposes of grounding moral responsibility. Their evaluative content should express the way a person conceives of a human life that is worth living and should have potential consequences as a criterion for choice and a guide for action. In other words, it should have a moral import. So evaluative beliefs about, say, the flavour of a particular dish or the performance of a certain car model are not of this sort, unless eating this particular dish or possessing this car are among the highest ends in a person's life (which, by the way, is not unthinkable nowadays). Given their role in the conduct of one's life, our real evaluative beliefs are likely to manifest themselves not only in our sincere general thoughts or declarations about our values but also in our particular judgements about concrete situations and in our actual choices and actions. But a straightforward logical behaviourism should be rejected, since there can be a real tension between our sincerely held evaluations and our actual behaviour, as well as between our deep, long-term values and our passing desires and preferences. Our particular choices and actions are not always a secure criterion of our evaluative beliefs.

If evaluative beliefs are to play a central role in grounding the possibility of ultimate regulative control over our choices and actions, and so of our moral responsibility for them, they have to satisfy a number of conditions: 1) Corresponding to the depth of moral responsibility ascriptions, they should be a central component of a person as a potentially morally responsible agent. 2) They should be correctly attributed to an agent as their true author and origin, in order for some of her choices and actions to be truly ascribable to her as their source. 3) The agent should have rational control over those beliefs; they should be justified and based on reasons. This requirement, however, should not give rise to a self-defeating infinite regress. 4) They should be potentially efficacious in our behaviour, even if we can sometimes act against them. 5) Corresponding to the regulative aspect of ultimate control, we should have alternatives with respect to them. 6)

The justification condition (condition 3) should hold even if the beliefs are not causally determined; in other words, our proposal should not fall prey to some version or other of the "Mind" argument.

In this section, we shall comment on the first four conditions, while leaving the two remaining conditions to the next two sections. However, we should warn that the conditions we have just listed are not fully independent of each other: what each condition involves should be understood in connection with the rest. So a general perspective of what ultimate regulative control amounts to in the context of our cognitive proposal is only to be expected if each aspect is viewed in the context provided by the others.

Let us start with the first condition. Throughout this book, we have insisted repeatedly that ascriptions of moral responsibility have a characteristic depth. In making such ascriptions, we are blaming or praising, and so valuing, the *person* for what she has done, as we view the action for which we hold *her* responsible as an expression of her self. If this is not the case, so that we see an action as a rather hasty and accidental event, and not really as an expression of some deep trait in the person, we tend to withdraw or significantly soften our judgement. Even Hume, whom Frankfurt rightly criticizes for having dealt only with freedom of action rather than freedom of the will, acknowledges this depth of moral responsibility ascriptions when he writes that "actions are objects of our moral sentiment, so far only as they are indications of the internal character, passions, and affections" (Hume 1975:99). And it is also the perception of this depth that has led thinkers, on both the compatibilist and the incompatibilist side, to look for the roots of moral responsibility in long-term features of the agent's self or character. Frankfurt's second-order volitions or Watson's valuational system are clear examples. Our proposal, centred on a subject's evaluative beliefs, honours this insight as well. Our evaluative beliefs are a central part of how we are, mentally speaking. They inform not only some particular decisions but also the general way in which we conduct our lives and our relations to other people. They are essential to the sense we can make of our life and of our place in the world. No psychological description of a person could be minimally complete that did not include her views about what she finds important and worth pursuing or avoiding in her life. And, if we reflect on our serious ascriptions of moral praise- and blameworthiness, we shall find ourselves ultimately praising or blaming an agent for her evaluative convictions. We seriously praise or blame someone for an action in so far as we see this action as a sign of, or as flowing from, her evaluative attitudes. So focusing on such attitudes as the ground of moral responsibility ascriptions does certainly account for the characteristic depth of such ascriptions as judgements about the person, her worth and value.

Let us move on to the second condition. According to the intuition of ultimate control as a requirement for moral responsibility, if we blame or praise a person for an action of hers in so far as we see this action as flowing from her evaluative views, then, for this judgement to be justified, it has to be the case that this person is the author of, and responsible for, these evaluative views. We consider the agent as praise- or blameworthy not only for her action, but also for her evaluative convictions, for we see her action as an expression of these convictions. As we pointed out in the preceding paragraph, if we do not, so that her action appears to us as expressing a rather momentary impulse, quite unrelated to her evaluative perspective, our judgement is significantly softened, or even

withdrawn, depending on several circumstances of the particular case. Our judgement of a person on account of her evaluative attitudes is not like our judgement of her on account of her intelligence or beauty. It seems that we are assuming that, unlike the latter properties, she has proper power or control over them, so that they are truly attributable to her. It is not, we think, that she has simply happened to have such attitudes, as she has happened to be intelligent or beautiful. We assume, then, that she is responsible for that part of her self that consists in such evaluations. In fact, if this assumption proves to be false, so that we cease to consider the person as the true origin of her evaluative views, as may well happen in CNC manipulation or Brave New World cases, we also stop viewing her as a proper target for moral responsibility ascriptions. But, then, under what conditions can a subject be held to be the author of her evaluative views? A look at the process through which we acquire such views will be useful in order to answer this question.

We acquire and shape our evaluative beliefs more or less in the same way as we acquire and shape our beliefs about matters of fact. Some natural, innate capacities are required in both cases. Especially important for forming correct evaluative beliefs, including moral beliefs, is the capacity to put oneself in the place of others, to try to see things from the perspective of fellow human beings. This seems essential in order to grasp the idea of a moral duty. Some subjects do not have this capacity, and this excludes them from moral responsibility ascriptions, given that they are not able to form correct evaluative moral beliefs. Moreover, as in the case of factual beliefs, we receive many of our initial evaluative beliefs from our social environment. We are initially told and taught which actions are good or bad, which ends are valuable or worthless, which attitudes towards oneself or others we ought or ought not to adopt, and so on. From the perspective we have now reached, this "external" origin of our evaluative beliefs should not lead us to conclude, as is done all too often, that we are not truly praise- or blameworthy for the evaluative beliefs we end up having, and for the actions that we perform in the light of them. If that conclusion were allowed, one should also hold, to go back to our example, that Newton is not truly praiseworthy for his theoretical work, since the sources of his first (and many of his later) beliefs about the physical world were also "external" to him. In fact, it is hard to see how we could have beliefs at all without external sources, including our social environment. Again, if human beings are constitutively social, it is as social beings that they can be morally responsible, morally praise- or blameworthy agents. The mere fact that we are constitutively social beings cannot be consistently used to deny our moral responsibility, for it is only as social beings that we can have moral duties and be morally responsible; but this does not imply that particular social conditions cannot undermine, destroy or promote, as the case may be, the moral responsibility of people subject to those conditions.

Now, if we really could do nothing but have the evaluative beliefs that we happen to acquire from our social environment, we would not be truly responsible for having them, and moral responsibility for our actions would not be possible. But we do not only acquire beliefs, either evaluative or merely factual; we also learn ways of forming, evaluating, accepting and rejecting them. In fact, in the absence of the latter, it is hard to see how someone could be said to have a system of beliefs. The most basic of these ways is related to the truth-aiming nature of belief itself. It consists in allowing our beliefs to be determined by how things actually are. We have referred to a higher phase of this

attitude, in relation to our examples in the preceding section, as respect and humility towards the structure of the things themselves. And we have also seen how this attitude of active passivity, as it might be described in a slightly paradoxical fashion, is not incompatible with true authorship and responsibility. Another central way of assessing beliefs that we acquire is to consider their logical relations. Contradiction is especially important in this respect: two contradictory beliefs cannot both be true. Again, this basic skill may be possessed in different degrees and is certainly perfectible. Let us stop here and refer to epistemology to get a more detailed view of cognitive assessment. Now, I think we use these basic ways of assessment not only with regard to our factual beliefs but also with regard to our evaluative ones. And this use is central to the possibility of moving on from a set of merely acquired evaluative beliefs to a system of evaluative beliefs that can be said to be the agent's own and for which she can be responsible. Though with a rather different aim in mind, I agree with Henry Richardson when, dealing with autonomy, he writes:

> Although Kant's contrast between acting on a desire and acting on principle is useful to understanding autonomy, he was wrong to think of the capacity of autonomy as existing within us a priori. Rather, it is a fragile and contingent achievement. It must, first of all, be achieved. Developing the capacity to act on a conception of reasons requires learning from experience. It requires being able to articulate one's reasons and subject them to testing.
>
> (Richardson 2001:296)

If Richardson's "testing" is enlarged so as to include not only empirical but also logical testing, his view comes quite close to our own. Not all views about evaluative facts that we receive from our social environment need to be consistent with one another, and this sets us the task of establishing which of two contradictory views is actually true. And, even in the absence of contradiction, we may also find that a received evaluative belief is not actually true: we may discover, from our own experience, that something is not really valuable and worthwhile even if we were told that it was.

Through the application of her capacity to assess evaluative views on the basis of logical criticism and experience (though with important qualifications to be made below about the sort of experience involved), a subject may progressively advance from purely received beliefs to a set of evaluative beliefs that can be said to be her own and for which she truly deserves moral praise or blame. In a rather modest way, as compared with such great intellectual achievements as Newton's *Principia Mathematica,* but not in a completely different sense, a system of evaluative beliefs can also be attributable to an agent as its author, and thus be a legitimate source of her praise- or blameworthiness. And, in so far as this system is truly hers, she can also be praised or blamed for acting in agreement or disagreement with it.

However, as we pointed out, a fuller account of this important condition of authorship, as a central component of the general requirement of ultimate regulative control on moral responsibility, is not independent of the other conditions we are dealing with, and especially of the two last conditions we listed, which will be dealt with in the next two sections of this chapter.

Something similar should be said about the third condition, which concerns the rational control and justification of our evaluative beliefs, with particular attention to the threat of an infinite regress: though we shall start commenting on this condition now, these comments should not be viewed in isolation. The fifth section of this chapter will be especially relevant to a proper understanding of it.

Let us first see whether the sort of infinite regress that makes true responsibility impossible, according to Strawson, threatens the rational control we might have over our evaluative beliefs. These beliefs are certainly a central component of "the way we are, mentally speaking", as Strawson puts it. But if Strawson's argument shows that we cannot be truly responsible for the way we are, mentally speaking, does it not also show that we cannot be truly responsible for our evaluative beliefs? Paraphrasing Strawson, if we are to have ultimate control over our actions, we must have ultimate control over our evaluative beliefs; this means that we must have chosen them rationally in the light of certain principles of choice (which include further evaluative beliefs), which in turn we must have chosen in the light of further principles, and so on. Consequently, ultimate control is impossible.

However, from a cognitive perspective on moral responsibility, this argument can be resisted. Strawson's argument assumes that the only basis of true responsibility (ultimate control) is rational choice. But this assumption is mistaken. Concerning a certain set of beliefs, it is not true that, in order to have ultimate control over them, that is, in order to be their true origin or author and to have rational control over them, one must have *chosen* that set in light of a further set that one must also have chosen in light of yet another set, and so on. Remember the example of the addition in the preceding section. The subject has arrived at a belief about the result of the addition by carefully applying the appropriate arithmetical rules. She is rationally justified in holding that belief, which she herself has given rise to. She has all the control over such a belief that is required for her being praiseworthy for having it. But she has not chosen the arithmetical rules by applying which she has reached her belief. It seems, then, that choosing the principles on which we form our beliefs is not required for having ultimate control over them. Similar considerations apply to the example of Newton. He did not choose the principles of inductive and deductive reasoning that he applied in his work, nor did he choose the empirical data that he worked with. In fact, as we argued, it is centrally important for his deserved praise that he did not choose either. And the same holds for the addition example.

It is the existence of a form of rational control which is not based on choice that places our proposal beyond the reach of Strawson's sceptical argument. Rational control over our beliefs has less to do with choice or voluntary production than with forming those beliefs in the right attitude of respect for how things actually are. And the same holds for our evaluative beliefs. We exercise rational and voluntary control over our actions by choosing them in the light of such beliefs, but the control we should have over those beliefs in order to have ultimate control over our actions does not derive from a further choice of those beliefs in the light of other beliefs. Instead, as in the case of non-evaluative beliefs, it has to do with appropriate respect for the facts they purport to capture, namely facts about what is really valuable and worth pursuing in human life. Just as it is important, in order to achieve our ends, that we have true, or at least justified, beliefs about matters of fact, in order to have a life worth living it is centrally important

that we have true or at least justified beliefs about which ends are really valuable and worth pursuing. Otherwise we may find that achieving our ends actually does nothing to promote our happiness and wellbeing.

However, even if our proposal is not threatened by a regress of choices, it may be affected by a different regress, a regress of reasons for believing. This is, however, part of a general epistemological problem, the problem of justification. It is a problem for everybody. But perhaps focusing on evaluative beliefs may throw some light on it. As we have suggested, experience plays a central role in the formation and assessment of our beliefs, both factual and evaluative. However, it is important to emphasize that the experience that is in play in evaluating, forming, modifying and rejecting evaluative beliefs, the experience that may convince us that a certain evaluative view is true or false and lead us to a set of such beliefs that can truly be said to be our own, is denser, and broader, than in the case of non-evaluative beliefs about matters of fact. It is closer to the sort of experience of life that some old people are said to enjoy. The term "wisdom" would seem to be more appropriate than "knowledge" to designate the kind of intellectual achievement that this important cognitive faculty is directed at. We have insisted on the central importance that having a correct set of evaluative beliefs has for our happiness, flourishing and wellbeing. In this connection, we must equally emphasize the importance of cultivating and developing the cognitive faculty we are referring to, for whether we end up having right evaluative views or not is strongly dependent on whether, and how, we exercise such a faculty.

Some remarks about this faculty whose aim we have called "wisdom" may cast some light on the justification of our evaluative beliefs, as well as on the way in which the problem of regress might be avoided in connection with them. Think of a basic true evaluative belief, namely that it is wrong to cause unnecessary suffering to innocent people, and compare it with a basic theoretical belief, such as that two plus three equals five. Though in both cases a sort of immediate seeing or direct intuition (in something like the way Descartes uses this term) is involved, there are some important differences in the quality of the intuition. Part of the difference has to do with the role that emotional states play in each case. The way we "see" that the evaluative practical belief is correct has to do with our imagining or thinking of actual cases that contravene it and the strong repugnance that they arouse in us. Unlike the case of the theoretical belief, our rejection of such contravening cases is not appropriately conveyed in connection with the term "absurd", but rather with the term "abject". It is, so to speak, a whole bodily rejection, not totally unlike our reaction to a putrid, ill-smelling piece of meat. We suggest that it is this holistic emotional involvement in our intuition of evaluative facts that actually prevents a regress of justification from arising in the case of evaluative beliefs. Concerning the evaluative belief we are referring to, someone's demand for a further justification could rightly be taken to reveal a cognitive and emotional impairment. To accept such a belief as simply true may actually be a requisite for someone to qualify as an agent whose evaluative moral views could be taken seriously, and so as a potentially morally responsible agent. In a different though related sense, it might be argued, as Christopher Hookway has actually done (cf. Hookway 2003), that affective states also play a central anti-sceptical, regress-stopping role in the case of factual and theoretical beliefs.

There are people who cannot actually see that certain basic evaluative propositions, such as the one that constitutes the object of the belief referred to, are simply true. They lack the faculty that leads to what we have called "wisdom" and are not able to reach justified evaluative beliefs. Though different evaluative beliefs may be true, and the assessment of truth and falsity in this field is bound to be complex, evaluative beliefs that are obviously false may call into question the quality of their holders as moral agents. A considered judgement about this matter, however, should also take into account whether an agent could have reached alternative, true beliefs. This will be the theme of the next section. However, we may now say that a severe lack of wisdom, in the sense indicated, excludes such alternative possibilities and therefore deprives the agent of the control that would be needed for her to count as a morally responsible agent.

Beyond these minimal thresholds, the importance of developing and perfecting wisdom has to do with the fact that there are many evaluative beliefs whose truth is much less obvious than the one we have considered, as well as with the fact that the application of even obviously true evaluative views to particular cases is not equally obvious itself. In some cases, it may be difficult to ascertain whether a particular situation is or is not actually a case of causing unnecessary suffering to innocent people, to use the same example. We shall come back to this point in the fifth section.

Let us finally address the fourth condition, namely that evaluative beliefs should be potentially efficacious in our behaviour. We have pointed out above that, if our cognitive proposal about the possibility of moral responsibility for our actions is to work, it is important that our evaluative beliefs be able to motivate us and give rise to effective desires to act in accordance to them. If we propose to ground ultimate control over our choices and actions in our evaluative beliefs, they should be able to motivate such choices and actions. Otherwise, even if we controlled our evaluative beliefs and were responsible for them, we could not thereby be responsible for our choices and actions. Now it is a central tenet of a Humean view of motivation that only desires, and not beliefs, are able to motivate. We have argued that, if this view is correct, ultimate control, and so moral responsibility in the sense of true desert, will be undermined, and we criticized Kane's apparent acceptance of a Humean view as jeopardizing his own libertarian perspective. As we pointed out with respect to the problem of justification, the question whether beliefs of an evaluative sort are able to motivate is a general issue with large ramifications, which extend far beyond the subject of this essay. However, we suggest that part of an affirmative answer to the question has much to do with the sort of cognitive access to evaluative facts that we have been talking about in the preceding paragraphs. Although there is discussion about the motivational potential of beliefs, it is largely agreed that emotional states can be motivationally effective. Now, it is *partly* because, and in so far as, we assess, correct and shape our evaluative beliefs through a denser and broader sort of experience than that related to matters of fact, one which involves our emotions, that the resulting beliefs can motivate us and give rise to desires to act in accordance with them. It is *partly* because, and in so far as, we judge that a certain evaluative belief is true or false on the basis of an experience that deeply involves our whole self, including our emotional system, that the resulting belief can have effects on our decisions and actions. Too narrow a view of both belief and emotion might be at the root of the widely shared Humean view that beliefs are not, as such, able to motivate an

agent's decisions and actions. However, a complete treatment of this question would have to be the subject of a different essay.

Let us go on to comment on the fifth condition, that is, the condition of alternative possibilities or regulative control, whose necessity for moral responsibility we have defended against several lines of attack.

Evaluative beliefs and alternative possibilities

We have extensively defended—especially in Chapter 2—an alternative possibilities requirement for moral responsibility against several attacks on it. We argued that, concerning moral responsibility for actions, we rightly take into account whether the agent had a choice, whether it was in her power to decide and to act otherwise. We have insisted that this requirement is related to the control we want to have over the moral responsibility we bear for what we do. Unless we have alternatives, we lose such control, for we may find ourselves morally responsible for something that we could not have avoided doing or bringing about. In the context of our cognitive proposal, the representation of alternative possibilities of action takes quite a standard form. Our evaluative beliefs form a complex system; they comprise more than our moral beliefs. We also think, for example, that our self-interest and wellbeing is worth promoting; and we sometimes face situations in which different evaluative beliefs, all of them our own, point to opposite directions and cannot be jointly honoured. Moreover, we also happen to have desires and urges that we may think are not worth promoting and acting on, but we may sometimes indulge in acting on them. This projects a rather classic picture of what having alternative possibilities of decision and action amounts to in connection with moral responsibility.

But if we also think that a deep, ultimate form of control over our actions is required for our moral responsibility, it seems that alternative possibilities should extend, as a component of such control, to the roots or springs of such decisions and actions as well. In the cognitive perspective we favour, moral responsibility is grounded in a subject's evaluative beliefs. We should try to show that we rightly take into account whether an agent could have had different evaluative beliefs in order to judge whether she is morally responsible, morally praise- or blameworthy, for having the beliefs that she actually has and, derivatively, for acting in the light of them.

However, we have rejected epistemic voluntarism, according to which we can have a belief just by choosing or willing to have it. In fact, part of the motivation of our proposal is to see whether scepticism about moral responsibility could be avoided, and we have suggested that such scepticism may arise out of a conative or volitive perspective on moral responsibility, as centrally grounded in choices or acts of will. So, as applied to our evaluative beliefs, the requirement of alternative possibilities should not be understood in terms of choice. We should not require that, in order to be morally responsible for her evaluative views, and for the actions that she performs in the light of them, a subject could have *chosen* to have different beliefs. This would be at odds both with our rejection of doxastic voluntarism and with the spirit of our proposal. In this context, we seem to have two options. We can hold that the alternative possibilities condition applies directly only to our decisions and actions, especially to those decisions and actions related to our

cognitive activity, and only indirectly and derivatively to our evaluative beliefs. Or we can hold that the requirement applies directly to our evaluative beliefs as well, and see what it means to say, in either case, that an agent could have had different evaluative beliefs. We can certainly interpret this locution in terms of the first option. On this interpretation, what we mean by it is that the subject could have carried out her cognitive activity in a different way, for instance by collecting more (or less) evidence or assessing more (or less) carefully the evidence she actually had. I think that this is what we sometimes mean by this locution. The question, however, is whether this is the only thing we mean or even could coherently mean by it. If the answer is affirmative, this gives the will a central role in the foundations of moral responsibility, which in turn, as we have tried to argue, might have harmful consequences for the prospects of avoiding scepticism. Taking this option, then, does not seem satisfactory. But I think that the locution can have a different meaning. We sometimes blame or praise someone for her moral views, and not just for the cognitive activity through which she reached them, which we may not know anything about. And in this case it seems that we are assuming that the subject could have had different moral views as a basis of our praise- or blameworthiness ascription. As a first clue to what that meaning may be, let us remember the importance we gave, for what concerns a non-volitive form of control over beliefs, to the attitude of respect and humility towards the real nature of the thing itself that one wants to have true beliefs about. But let us try to get a more intuitive sense of what that meaning may be by considering some examples where the question whether an agent could have had different evaluative beliefs plays a central role in the moral responsibility she bears.

Think first of a landowner in Ancient Greece, 4th century BC, who buys a slave to work on his property. Suppose that the landowner, call him Meno, is psychologically and cognitively normal. Let us also assume that buying slaves is a morally wrong action. On these assumptions, is Meno morally blameworthy for doing so?

Let us see what the judgement about this case would be from the perspectives of compatibilism and our own proposal.

Think of compatibilism. Suppose that determinism is true. It is plausible to assume that the condition of control, on a compatibilist construal of it, is met by Meno: he bought the slave because he decided to, and his decision was made for appropriate reasons. He also satisfies a conditional, compatibilist construal of the alternative possibilities condition: if he had decided not to buy the slave, he would not have bought him. He was also appropriately responsive to reasons (Fischer): if he had faced important reasons for not buying the slave, he would have recognized those reasons and would have decided and acted accordingly. It seems, then, that, on the correct assumption that buying slaves is morally wrong, the verdict of compatibilism should be that Meno was morally responsible, morally blameworthy in fact, for buying the slave.

From our cognitive perspective, however, the verdict about this case differs. Meno, the landowner, did something morally wrong in buying the slave, but he is not morally blameworthy for doing so. We may assume that he was the author of his evaluative beliefs, which included the belief that buying slaves is not morally wrong. We may also assume that he had alternative possibilities of decision and action, in a conditional and even in a categorical sense. But he did not have relevant alternative possibilities concerning his evaluative belief that buying slaves is not morally wrong. His belief was false and, in acting on it, he was doing something morally wrong. But he was not morally

blameworthy for his action, because, given his historical and social circumstances, the true belief that buying slaves is morally wrong was not within his possible cognitive landscape; it was not available to him. Even if Meno had formed his evaluative beliefs with the right attitude of humility and respect for evaluative facts, it is not reasonable to expect that he could have seen that truth, which started to emerge and spread only four centuries later, with Christianity. Even though Meno can be said to have control over his evaluative belief that buying slaves is morally permissible, and to be its true author, as these concepts were sketched out in the preceding sections, he did not have ultimate *regulative* control over that belief: *he could not have had a relevantly different belief,* namely the belief that buying slaves is morally wrong. And this is why the judgement should be that Meno is not morally blameworthy for buying the slave. But I think that this is also the judgement that our natural, pre-theoretical attitude towards this case will yield.

A consequence of our perspective that can be seen in the discussion of this example is that, even if moral objectivism is true, so that the truth of moral propositions is not relative to particular circumstances and times, moral responsibility, instead, is relative to circumstances and times. To see this, think now of another character, Timothy, a landowner in mid-19th-century South Carolina, who also buys a slave who can work on his property. Assume, as before, that buying slaves is morally wrong and that Timothy was also psychologically and cognitively normal. Suppose also that he thinks, like his Ancient Greek counterpart, that buying slaves is not morally wrong. Timothy thinks so, we may suppose, owing to his upbringing in a slave state and environment. Now, though an accurate judgement about his moral responsibility would require filling in more details, I think that, with the elements at hand, the correct verdict in this case is that he is morally blameworthy for buying the slave. He does not think he is doing something morally wrong, but in this case he could, and should, have had a different belief. The truth that slavery is a morally wrong institution, and so that buying a slave is not morally permissible, was within his possible cognitive landscape. He, unlike Meno, his Greek counterpart, could have believed that truth. For Meno, this moral truth lay four centuries ahead, but for Timothy it was eighteen centuries behind. Moreover, he lived in a Christian environment, in which such a truth could be found relatively easily. And, for an educated man, the arguments against it, concerning for example the inferiority of black people, could easily be seen as fallacious. Therefore Timothy is morally blameworthy for buying the slave, given that this action comes from a belief over which he had ultimate regulative control.

In the light of all this, let us see what having direct regulative control over one's beliefs may consist in. In saying that Timothy, the American landowner, could have had a different evaluative belief we are not simply saying that he could have carried out an appropriate and more careful cognitive activity, though we may also mean this. We are also saying that he could (and ought to) have *seen* what was there to be seen. He was not humble and respectful enough towards patent evaluative facts. There are some connections between this judgement and cases in which we blame someone for something that she involuntarily did or omitted. Our partner may blame us for forgetting her birthday or for forgetting an appointment for dinner we had agreed. Forgetting is obviously not voluntary. It is not even an action. Our omission is not voluntary either. But I think this ascription of blame is not like blaming someone for, say, being ugly.

The latter judgement is not fair, but it seems to me that the former can be. And what explains the difference is the assumption that we have a form of control over our forgetting that we do not have over our physical constitution. In some sense of the words, it is true that we ought not to have forgotten our partner's birthday or the appointment we had for dinner, and it is also true, in some sense of the words, that we could have remembered both things. When we say that we could have remembered such things, we are not merely saying that it was logically possible that we should remember them. And when we say that we ought to have remembered them, we are speaking of moral obligation. These "could" and "ought" are those that feature in the "ought"-implies-"can" principle and in the alternative possibilities condition of moral responsibility. We are saying that remembering those things was our duty and that it was within our reach or power. And the regulative control that is being assumed in the corresponding blame ascriptions is, I think, close to the regulative control we assume people sometimes have over their beliefs, including their evaluative beliefs. It is the sort of non-volitive regulative control that we have been suggesting.

Even if believing, like forgetting, is not voluntary, and not an action, we can justifiably blame someone for her beliefs, as our partner can justifiably blame us for forgetting our appointment. The analogy goes quite far, which suggests that we are facing closely related phenomena. In blaming us for forgetting our appointment, our partner is blaming us for our blindness to both our appointment and her, which reveals our lack of respect for both. Similarly, in blaming the American landowner for his belief that slavery was morally permissible, we are blaming him for *not seeing* what was there to be seen, namely the moral evil in slavery and the slaves themselves as his fellow human beings, which in turn reveals his lack of respect for them.

If we agree that it makes sense and can be justified to ascribe moral blame to someone for forgetting an appointment (and not, say, for her ugliness) and to assume, as a basis for that judgement, that she ought to, and could, have remembered it, then we are acknowledging that there is a form of regulative control that does not depend on choice and that is relevant to moral responsibility. It is plausible to think that we have this sort of control over our beliefs and that we assume that persons have it when we morally blame (or praise) them for their moral and evaluative beliefs, and not only for their actions. It would be important to go from the rough idea of this sort of control which we have derived from the examples to an investigation of its precise structure and conditions. But, on the basis of the preceding examples, we have reason to think that it exists. And, if our diagnosis of the roots of scepticism about moral responsibility is correct, its existence may be significant for the prospects of overcoming the sceptical predicament.

Seeing something that is there is not an action, but it is something for which one can sometimes deserve merit, as is the exercise of a capacity that can be cultivated and perfected. This is especially true with respect to evaluative facts. Respect for others is a basic moral insight that most people get from their social environment, and it provides a reason for developing that capacity and improving our receptiveness and sensitivity towards the needs and value of other human beings. Many circumstances can and should be taken into account in evaluating an agent's praise- or blameworthiness for seeing, or not seeing, certain evaluative facts, and for having the corresponding beliefs. More favourable circumstances may lead to a more rigorous judgement about an agent if she does not develop right evaluative views, and less favourable circumstances may assuage

or even preclude an agent's blame for not doing so. Conversely, an agent who develops her sensitivity to evaluative facts in a difficult environment may deserve more praise than someone raised in a more hospitable situation.

Though our examples have dealt with blameworthiness, alternative possibilities are also relevant to an agent's praiseworthiness for her evaluative views. This assertion goes against Wolf's asymmetry thesis, according to which alternative possibilities are required for an agent's blameworthiness, but not for her praiseworthiness. In Chapter 2, we argued for the general application of the alternative possibilities requirement with respect to actions. And we think that something similar can be said about evaluative beliefs. Just as an agent deserves blame for her wrong evaluative views in so far as it was in her power to have formed different, and better, views, an agent deserves praise for her correct evaluative views in so far as she could have formed a worse set of such views. These judgements should be qualified on account of several factors, including the circumstances in which she grew up and her natural gifts, but it seems to me that the consideration of those alternative possibilities informs and underlies such judgements.

It may seem that, in requiring the availability of alternatives for an agent's praiseworthiness for her actions and evaluative beliefs, we are assigning a positive value to a kind of freedom, namely the freedom to act wrongly or to form mistaken moral beliefs, which, in Dennett's expression, would not be "worth wanting". But I think this compatibilist move is wrong. In requiring alternatives in order for an agent to be praiseworthy for her good actions or correct evaluative views, we are not giving a positive value to the possibility of acting in a bad way or forming wrong evaluative views. We give a negative assessment of this possibility, but, at least for human beings, the possibility is there, and it is in contrast with it that we praise an agent for not taking it. In a related vein, Wolf (1990:81–2) rejects a consequence of such a requirement, namely that a person whose moral convictions and commitments are so strong that she is literally incapable of acting in a morally wrong way is not praiseworthy for her good actions. This consequence, however, is not obviously false. For us human beings, there is pressure, coming from some of our desires and natural inclinations, towards morally wrong ways of acting and believing. In the absence of such a pressure, in the absence of any countervailing factors to our invariable disposition to form objectively correct moral beliefs and act on them, in the absence of any temptation, we would be good persons but we might not actually deserve praise for those beliefs and actions. Just as a person can perform a morally bad action and not be blameworthy for doing so (remember our Ancient Greek landowner), a person can perform a morally good action and not be praiseworthy for doing so either. And the reason, in both cases, is that they lacked regulative control: they could not have believed or acted otherwise. This is partly why we judge that subjects in a Platonic New World, who are determined to have right values and to act on them, are not morally praiseworthy, even if they are good people. True authorship, as a basis for true desert, requires the availability of alternatives.

Whether a person has deep regulative control over her choices and actions, and so is morally responsible for them, is difficult to establish. This means that our moral responsibility ascriptions may be wrong in various particular cases. But in focusing on evaluative beliefs, rather than choices and willings, as the root of moral responsibility, we are trying to show that moral responsibility is indeed possible, even if its presence and degree may be hard to assess in many or even most cases. A crucial test for our

incompatibilist, libertarian proposal is whether it can meet the classical compatibilist objection against libertarianism, according to which indeterminism undermines the control that would be required for moral responsibility. Remember that meeting this objection is the sixth and last condition we have required for the plausibility and anti-sceptical effectiveness of our cognitive view of moral responsibility. The "Mind" argument, in its different versions, is the standard formulation of this objection. Let us now address this complex and important question.

Indeterminism, belief and moral responsibility

The purpose of this section is to see whether indeterminism can be reconciled with rational control. We shall distinguish two perspectives on the place and role of indeterminism in practical rationality and contend that, though one of them is unlikely to provide a satisfactory response to worries about rational control in indeterministic contexts, the other might well be able to do so. On this basis, we shall be arguing that a cognitive approach to moral responsibility has better chances than a conative, will-centred approach of giving a justified positive answer to that question.

Many libertarians see a metaphysical basis of free will and moral responsibility in quantum indeterminism, especially as it may take place in the human brain. Kane, as we have seen, is among them. If quantum indeterminism at the subatomic level were amplified, instead of cancelled, at higher levels of the organization of matter, such as the neuronal level, this would make the brain into an indeterministic system, so allowing for alternative neural pathways and, on appropriate views of the relationship between brain and mind, for the alternative choices and actions that, according to libertarians, including the author of this book, are required for moral responsibility.

Now, though we accept the alternative possibilities requirement, we do not share those libertarians' hopes about the perspective we are considering. On this perspective, indeterminism at the subatomic level would be amplified and transmitted to higher neurological and mental processes through metaphysical relations such as supervenience or even identity. We can call this view of the role of indeterminism in mental acts and processes "bottom-up" indeterminism. We do not see how to avoid the conclusion that, even if it provides room for alternative pathways of decision and action, bottom-up indeterminism sweeps away the agent's rational and volitional control over them, so eroding the basis of moral responsibility. Quantum phenomena are among the most basic physical processes. As we pointed out in the preceding chapter, if quantum changes suffered by a subatomic particle are actually indeterministic, they are not under anyones's control. So any attempt to ground free will and moral responsibility in such basic indeterministic processes will stumble, at some point, on the problem of the agent's control over her choices and actions. It would seem, then, that the versions of libertarianism that accept bottom-up indeterminism as a metaphysical basis for free will and moral responsibility are bound to fall prey to some construal or other of the "Mind" argument, whose last consequence is SMR's premise C, namely the impossibility of moral responsibility given indeterminism. The fact that a quantum-cum-chaotic process occurs in my brain which is identical or subvenient to a certain decision does not make that process and decision truly my own, in the sense that would be needed for moral

responsibility, and neither does it allow me to control them rationally. Both aspects of ultimate control, namely ultimacy of source and rational control, are actually undermined by bottom-up indeterminism.

If indeterminism is to make free will and moral responsibility possible, it has to hold at the right places and have the right relations to the mental and neurological phenomena, so as to provide room for rational control. We propose to call an indeterminism of this sort "top-down" indeterminism. According to this perspective, the sort of indeterminism that can allow for rational control holds primarily at "higher" levels of reality, especially at the level of rule-governed, normative systems, practices and institutions, such as language itself, as well as written or unwritten codes regulating various aspects of social interaction within human societies. These rule-governed practices include practical deliberation. What has to be shown is that, at this level, indeterminism holds and is none the less compatible with rational control over alternative pathways. As a second step, we should tell a reasonable story about how indeterminism at this higher level could be transmitted downwards, so as to contribute to the shaping of the human brain and some of the neurological processes that take place therein. As the reader will have noticed, this second step involves an attempt to deal with the problem of mental causation, which is an aspect (the possibility of mind-to-body causal influence) of the traditional mind—body problem. Our aim concerning this venerable and recalcitrant problem will be quite modest. We shall try to show that downward causal influence, from mental, content-bearing states or events to physical ones, is neither unintelligible nor necessarily at odds with a scientific outlook.

Let us go on to the first task. One sense in which normative systems, practices and institutions can be said to be indeterministic is that, as Peter Winch (1963), following Wittgenstein, insisted, the distinction between correct and incorrect ways of acting in relation to the corresponding normative standards is constitutive of them. But this distinction involves the notion of alternative possibilities. They can be said to be indeterministic in a second and important sense, namely that, in several particular cases, the system's normative standards do not determine one single way of complying with (or breaking) them and of acting in a correct (or incorrect) way; in relation to this, in many cases there is wide room for reasoned argument and discussion about whether a particular way of acting is or is not correct with respect to the system's normative demands. This is why legal codes, for example, essentially need specific institutions to apply them to particular cases. This is not due to the fact that the codes are not sufficiently precise; it is rather within the nature of a rule or norm that the question of what it means to comply with it can always arise in certain cases. So, in the administration of justice, reasoned verdicts about particular cases (jurisprudence) become an integral part of the normative criteria for further judgements. In relation to this, the question whether a particular act is or is not justified with respect to certain normative criteria can sensibly be raised on many occasions. The concept of justification is, then, essentially open-ended and relative to circumstances of several kinds. Finally, normative systems and practices are indeterministic in the sense that there is room for reasoned discussion about whether a particular situation is one to which a certain normative judgement applies. To take a previous example, even if one accepts the principle that it is morally wrong to cause unnecessary suffering to innocent people, this does not decide the question whether a particular situation actually falls within that principle's scope.

The significance that this reference to normative systems has for our proposal about the possibility of moral responsibility partly has to do with the fact that our evaluative (and non-evaluative) beliefs also form a system that is subject to normative standards related to the constitutive aim of truth. Demands for coherence or responsiveness to the relevant evidence bear upon human beings as subjects of beliefs. Our belief systems are indeterministic in the three senses we have distinguished in the preceding paragraph. The distinction between correct and incorrect, justified and unjustified beliefs and ways of forming them is constitutive to the possession of such a system. There is also room for discussion about whether and when a certain belief or belief set complies with the normative demands and so is actually correct or justified. This second feature, according to which, in many particular cases, the normative standards do not determine what it is to comply with them and so whether a certain belief is correct or justified, makes it the case that the holding of particular beliefs in certain paradigmatic or extreme cases often acts as a negative touchstone of the ability for such a compliance. As we saw in the preceding section, rejection of certain basic evaluative beliefs may disqualify a person as a morally responsible agent, for we take this rejection to be a clear sign of her inability to conform to normative standards and form correct moral beliefs. Our previous example of someone's seriously holding that there is nothing morally wrong in causing unnecessary suffering to innocent people may be a case of this kind. Finally, it also seems clear that there is room for reasoned discussion about whether a certain belief is or is not a case to which a doxastic normative standard applies.

It is not only our system of evaluative beliefs that is subject to normative standards. Our practical deliberation, by which we try to evaluate possible ways of acting and reach decisions in the light of those assessments, is a practice that is also subject to normative demands. As Christopher Hookway writes, "reflection, including both practical and theoretical deliberation, is an activity that can be controlled through the exercise of normative standards" (Hookway 2001:190). Central among those standards are our evaluative beliefs about the worthiness of several ends and purposes. And just as, concerning our evaluative beliefs, there is room for reasoned discussion about whether a particular belief does comply with the normative standards of coherence or responsiveness to facts and experience, or about whether a particular belief falls within the scope of a standard, there is a similar free space concerning our practical judgements and decisions for what concerns their relationship to our evaluative beliefs. For example, for an agent who highly values friendship it is still an open question whether a certain practical judgement or a decision about a particular way of acting is actually one that honours that belief or is the one that best honours it. And it may also be an open question whether a certain way of acting violates that evaluative belief.

Thus we have at least two spheres closely related to moral responsibility (our evaluative beliefs and our practical deliberation) to which normative standards constitutively apply, and each of them shows at least the three forms of indeterminism that we have indicated concerning normative systems and practices in general. Moreover, in spite of the hierarchical ordering between these two spheres, practical deliberation can also retrospectively affect our evaluative beliefs. For example, a decision against one of our evaluative beliefs, perhaps prompted by a momentary impulse or special emotional state, even if it is presently experienced as wrong and accompanied by a feeling of guilt, may end up showing us that it was the belief, not the decision or the corresponding

action, that was actually incorrect. And, given the inner dependence relations among our evaluative beliefs, this can lead to a reorganization of our whole system of such beliefs or a part thereof.

The complex net of interactions between these two normatively guided spheres, each with its own field of alternative possibilities, opens up an even wider indeterministic field of free movement for us as believers and practical deliberators. However, it is centrally important to realize that this indeterminism related to normative systems and practices, unlike the indeterminism related to quantum phenomena, is essentially linked to reasons and reasoned discussion and criticism, and thereby initially receptive to rational control. It is the sort of indeterminism that, being essentially associated with the possibility of rational argument, may give rise to the sort of non-arbitrary, reason-related freedom that can ground moral responsibility. This is the sort of freedom that grows in what John McDowell, following Sellars, has aptly called "the space of reasons". As he writes, "this freedom, exemplified in responsible acts of judging, is essentially a matter of being answerable to criticism in the light of rationally relevant considerations" (McDowell 1998:434).

This wide field of possibilities of belief, judgement and decision, backed by reasons, also opens up the space for forging rationalizations about the true motives of our choices and actions, inventing excuses for what may appear as morally unjustified and blameworthy actions, and even for self-deception about our actual reasons or motives. These rationally flawed phenomena pay retrospective tribute to reason, however, in that they are only possible and acceptable in so far as they present themselves in the guise of true and rationally justified explanations.

Moreover, normative systems are significant to the possibility of moral responsibility, not only in connection with the rationality aspect of ultimate control but also with regard to its ultimacy of source aspect. We have argued that, in appropriate conditions, we can truly be considered as authors of our evaluative beliefs, even if they trace back to factors external to ourselves. We may now see how external factors can actually be indispensable to our status as subjects of evaluative beliefs, if responsiveness to normative standards is constitutive of such a status. For it is only through socialization and education in a human society that we can become participants in normative systems and practices. And it is as participants in such systems and practices that the idea of alternative possibilities and of reasons for each of them, as well as the idea of rational justification and control over our beliefs and choices, make sense to us. It is as social beings, as members of objective systems governed by normative standards, that we become capable of rational control, deliberation and reflection. And, as these abilities are a condition of someone's qualifying as a morally responsible agent, it is as social beings that we can be agents to whom moral responsibility ascriptions can justifiably apply. As we have insisted in previous sections, the social nature of human beings should be viewed as an enabling condition for their moral responsibility rather than as an obstacle to it. We can now give further support to that contention.

Let us now move on to the second step in the analysis of top-down indeterminism as an enabling ground of moral responsibility. Even if normative systems and practices can rightly be taken to be, in McDowell's words, "second nature" to us, it is important to see whether the reason-related indeterminism that we have found in them could reasonably be taken to transmit itself downwards, so as to partially shape our "first nature" as natural

biological systems. Otherwise, if reason-insensitive determinism or quantum indeterminism reigns unopposed at the neurological level, unaffected by the reason-sensitive freedom of the "space of reasons", our proposal might fall prey to the objections of scientific obscurantism traditionally levelled against libertarianism, or be open to the charge of erecting a problematic form of dualism, on the basis of an unrestricted opposition between freedom and nature. And these objections, in turn, might prompt the sceptical suspicion that belief in freedom and moral responsibility, psychologically inevitable as it may be, is none the less an illusion.

Let us think of some examples of normative systems and practices, such as musical notation and its interpretation with a musical instrument, or arithmetic and its application by a shop assistant to calculate the amount owed by customers. In both cases there is a body of theory with an objective content and rules. And anyone wanting to master those systems should be prepared to learn that content and be trained in applying the rules. Both cases show, in different degrees, the levels of indeterminacy we distinguished when we referred to normative systems in general. Now think of learning to play a musical instrument, say the classical guitar, after having being trained in reading musical notation. One has to translate musical signs into sounds by pressing the strings with one's fingertips on the right places of the guitar neck while simultaneously plucking the right strings with the fingertips and nails of the other hand. In the first stages of this learning process, the movements are clumsy and difficult to perform; one has to think of the steps involved in interpreting each note. Gradually, however, through a painful process and a good deal of practice, the movements of the fingers become less reflective and more automatic; the reading and translating of the musical notation into sounds is less of a problem, so that one can then concentrate on other aspects of the interpretation, such as expressiveness or the clarity of the sound; aesthetic pleasure gradually increases and playing becomes more rewarding. As one confronts a difficult new piece, the finger movements are not as sure as in pieces one has already studied, but, also gradually, the process of managing to play it decently becomes shorter and less painful than on prior occasions.

Let us stop here in this (not too inexact, I hope) phenomenological description of the process of learning to play the guitar. Now the question is what happens during this learning process at the neurological level. Here we must leave the firmer ground of description and proceed to the shaky terrain of speculation. But we do not aspire to empirical accuracy. It will be enough if we are able to suggest a reasonable and fairly general hypothesis not at odds with plausible neuroscience. Here it is. Corresponding to the progress in the learning, what takes place in the brain and the nervous system is the establishment of an increasingly complex system of neural connections, which include afferent and efferent nerve fibres. These connections are reinforced with practice, which explains the increasing sureness and automatic character of one's movements during the process of learning, and they are also correspondingly weakened by lack or insufficiency of that practice, which in turn explains why, after a long period with no playing, the performance becomes poorer and the movements clumsier. If the period is very long, the subject can even find herself no longer able to play pieces she used to play quite easily in the past. In the end, she may find that she needs to go through the learning process again almost from the start. This second process, however, is likely to be shorter than the first,

owing to the remaining neural connections. In a nutshell, at the neurological level the process of learning consists in the forming of a certain configuration of neural networks.

We can now go back to philosophy. It is the neural configuration that, at least partially, explains someone's ability to play the guitar—which is a particular normative practice—with a certain degree of decency. However, what at least partially explains the fact that this neural configuration, with its various alternative pathways and complexities, has been formed, as well as the form it actually has, is the objective content of normative systems, institutions and practices related to music and guitar playing. Through the process of learning, the several indeterministic levels involved in those systems and practices have come to shape the structure of the brain and nervous system of the learner. New forked neural pathways have been shaped through learning, according to the alternative pathways corresponding to the levels of indeterminism of the normative systems and practices involved in the learning. And, since the learning never actually ends, this indeterministic shaping of the brain goes on with it, becoming more complex and subtle, and more sensitive to tiny variations in neural activation levels. However, since this branching neural structure has been shaped, through the learning process, according to the reason-sensitive alternative possibilities constitutive of the normative systems and practices involved, it does not undermine the agent's rational control over those alternatives, but rather enables and enhances it. At the same time, however, some physical and biochemical processes in the brain unrelated to reasons, such as natural decay or death of neuronal tissues, hormonal insufficiency or even indeterministic changes at the quantum level, act in the opposite direction, increasing entropy and damaging rational control.

Note that this (admittedly rough) picture of the relationships between the "space of reasons" and the brain, between mind and body, to put it in more traditional terms, is compatible with supervenience of mental properties on physical ones. Think of an agent in a certain phase of her process of learning to play the guitar and imagine that an exact physical duplicate of her suddenly comes or is brought into existence. Given supervenience, this creature will also be able to play the guitar with the same level of perfection as the original agent. But some comments are in order. First, I take it that the probabilities of such a physical duplicate coming to exist without herself going through a process of learning are astronomically low. Moreover, unless the new creature engages in a process of learning to play the guitar, so that her inborn abilities are cultivated and perfected, she will quickly cease to be a physical-cum-mental duplicate of the original agent. Finally, though the original agent is truly praiseworthy for her ability to play the guitar with a certain degree of perfection, her physical duplicate is not. This, however, does not deny supervenience. It is rather the case that some properties, both mental and physical, are historical and dependent on their origin and the actual causal interactions that gave rise to them, so that they supervene on a broader physical basis. Being a sunburn, to take a Davidsonian example, is one of them. An alteration of the skin physically identical to a real sunburn but not caused by the sun is not itself a sunburn, because it does not have the right origin. Being a planet, a Fodorian example, is another. As for mental properties, some of them also show this feature. Remembering something is one example. Being praise- or blameworthy (including moral praise and blame) for a certain action or achievement is another. Now, if the physical duplicate of our musical heroine had come to her present physical state, which enables her to play the guitar,

through the same physical history and causal interactions as the original agent, she would also be praiseworthy for her present ability and performances, for she would also have gone through a learning process and those achievements would rightly be attributable to her. If, however, she had been suddenly brought into existence, as we initially assumed, she would not deserve praise for those abilities and performances. This, incidentally, is at least part of the reason why Brave New World citizens, provided that they have come to have their evaluative views through brain manipulation, are not morally responsible, praise- or blameworthy, agents.

With due respect for the differences, the preceding remarks can also be applied to many other normative systems and practices, including those directly related to moral responsibility, such as an agent's system of evaluative beliefs and her practical deliberation.

Now, if top-down indeterminism, as we have characterized it, actually holds, that is, if, as we have argued, normative systems and practices constitutively incorporate alternative pathways backed by reasons at the different levels we distinguished, and if this reason-sensitive indeterminism can actually transmit itself downwards so as to partially shape the inner structure of our brains, then, provided that our evaluative beliefs and practical deliberation are normative in that sense, we can have the makings of a response to the traditional compatibilist objection to incompatibilism, namely that indeterminism precludes rational control over our decisions and actions and, with it, moral responsibility itself. Let us see how this worry might be dispelled.

We shall contend that the objection is powerful, perhaps decisive, against traditional, will- or choice-centred versions of libertarianism, especially when combined with what we have called a bottom-up conception of indeterminism, but that a cognitive libertarian approach of the kind we have suggested might well be able to meet it. We shall address this worry (the "Mind" argument) with the aid of several versions of it that we went through in the first section of Chapter 4.

Remember Van Inwagen's example of the judge (J) who, after calm and careful deliberation, decides not to raise his hand at time T and so not to grant special clemency to a convict. J does not raise his hand at T and the convict is sentenced to death. As we have seen, we are indebted to Mele for one powerful way of formulating the "Mind" argument on the basis of examples of this kind. Adapting Mele's suggestion to Van Inwagen's example, imagine a close possible world, with the same past and natural laws as the world J inhabits, in which a twin of J's, call him J*, exists. Suppose now that the only difference between these two worlds comes at time T. At that time, while J decides not to raise his hand, J* decides to raise his. Remember Mele's words: "If…there is nothing about the agents' powers, capacities, states of mind, moral character, and the like that explains this difference in outcome, then the difference really is just a matter of luck" (Mele 1999:99).

We do not see how to avoid Mele's conclusion if we hold, as libertarians traditionally do, that, for moral responsibility to be possible, a different choice has to be compatible with the same past and natural laws that led to the agent's actual choice. This contention is a form of the categorical interpretation of the alternative possibilities condition on moral responsibility. To see what is wrong with this understanding of the condition, let us go back to the example of J. As Van Inwagen presents the example, J goes through "a suitable period of calm, rational, and relevant deliberation" (Van Inwagen 1983:69). So

we can plausibly assume that, at a certain time in his deliberation, J arrives at a practical judgement according to which, all things considered, it is best to deny clemency to the convict, and so not to raise his hand, or at least better than the relevant alternative. Now, J* also arrives at this practical judgement and, unlike J, decides to raise his hand. Given that J and J* are perfect twins until the time of the choice, this presents J's choice as purely chancy and arbitrary. He might equally have decided to raise his hand. So, if we insist that the choice is undetermined until the very moment it is actually made, we are severing the link between practical judgement and choice. We are accepting that, for moral responsibility to be possible, an agent must be able to decide to do B even though she judges that doing A is better. Decisions of this sort may happen if, for example, cases of weakness of the will do actually occur, but weakness of the will is, in everybody's lights, an irrational phenomenon. To insist on this sort of movement, then, amounts to grounding moral responsibility in arbitrary or irrational choices and paving the way for the compatibilist objection. Indeterminism placed at this particular juncture, that is, between practical judgement and decision, certainly erodes the agent's rational control over her choices. To resist Mele's objection by holding that, in quantum-cum-chaotic scenarios, sameness of individuals or their stories is not defined, as Kane does, actually worsens things, for that move strongly suggests that a bottom-up conception of indeterminism is being assumed, and with it the dependence of choices on quantum changes in the brain, beyond the agent's rational or even volitional control.

At least as powerful as this version of the objection is the "Rolling-Back" version. In Van Inwagen's presentation it features an agent, Alice, who, in a difficult situation, confronts a choice between lying and telling the truth, and freely chooses to tell the truth. Imagine that God reverts the world to the state it was in before that choice and allows it to go "forward" again. Supposing that her act is undetermined, we can plausibly assume that, after a high number of replays, the ratio of each possible outcome with respect to the total number of replays converges on some value, say 0.5. On this basis, the only reasonable option is, for each new replay, to assign the same probability, 0.5, to Alice's lying and to her telling the truth. So this was also the probability of Alice's lying, and of her telling the truth, in the actual, initial situation. The conclusion, in Van Inwagen's words, is that "an undetermined action is simply a matter of chance" (Van Inwagen 2000:15), as compatibilists have traditionally held. The argument works equally well for choices or decisions, instead of acts. The problem would seem to be roughly the same as in Mele's version of the objection.

Now, how does our proposal fare with respect to these objections? Corresponding to the cognitive character of our approach to moral responsibility, and contrary to volitive approaches, the main role in the process of practical deliberation should be assigned to practical judgements rather than choices. In the picture we have sketched above, practical judgements about which action is best, or better than the alternatives, should be understood as the application of an agent's evaluative beliefs, as normative standards, to a particular situation she faces. A practical judgement is a cognitive rather than a volitive phenomenon. It is not properly a choice, but, so to speak, the expression of a belief about the value of particular ways of acting, formed in the light of more general evaluative views. As a rule, an agent's rational choices are fixed by her practical judgements. Choices cannot ride free from practical judgements if they are not to fall into arbitrariness. On our cognitive approach, then, practical judgements replace choices as

the central steps in practical deliberation. Choices, in turn, are dependent on practical judgements and remain linked to them. At a deeper level, which grounds the agent's ultimate control over her choices and actions, we have again, of course, beliefs—of an evaluative sort—rather than radical choices (e.g., Kane's self-forming willings). These beliefs are formed, in turn, in the light of another set of normative standards, such as coherence and sensitivity to the facts.

This picture is not a mere reorganization of the components of practical deliberation. Giving centre stage to practical judgements, conceived as applications of more general evaluative views to particular cases, allows retaining indeterminism in practical reasoning without giving up rational control. As we pointed out above, application of a normative standard to a particular case is not a mechanical task. It leaves space open for rational discussion about both whether a particular way of proceeding complies with the standard and whether the particular case at hand actually falls within the scope of the standard. It is, then, both an indeterministic process and a non-arbitrary, reason-sensitive one. In this context, we need not deny the libertarian view that choices are undetermined; but we view the indeterminacy of choices as derived from the reason-sensitive indeterminacy of practical judgements, so protecting choices themselves against charges of arbitrariness or irrationality.

With these considerations in place, let us return to the objections and see whether we can plausibly resist them.

Think first of Mele's "twins" objection, which we applied to the judge example devised by Van Inwagen. We may assume, as Van Inwagen also does, that J is rationally and emotionally normal, and capable of competent practical reasoning. We may then assume that, at a certain moment in his "calm, rational, and relevant" practical deliberation, he arrived at the practical judgement that sentencing the convict to death, by not raising his hand, was better than the relevant alternative. Assuming rationality again, we may also accept that, at a deeper level, he holds the belief that capital punishment is justified, at least for some especially serious crimes. In these circumstances, his decision not to raise his hand and his corresponding action are ones he has rational control over. (An additional question is whether he has ultimate regulative control over this belief, and so whether he is morally responsible for his practical judgement, choice and action. This depends on his having met the conditions of true authorship with respect to that belief, including the availability of the opposite—and, in my view, true—belief that capital punishment is not justified.) Now imagine J's psychological twin, J*. The objection is supposed to be that, if choice is causally undetermined, J* may well make a different decision than J, namely to grant clemency by raising his hand, which shows that, contrary to the hypothesis, J's original choice was a matter of chance and so beyond his rational control.

Now, are we forced to accept this conclusion? In order to answer this question, we have to ask to what extent the sameness of both twins and their histories is supposed to hold. If they and their respective histories were exactly alike until the very moment of choice, *including the practical judgement that not granting clemency is better than the opposite,* then the conclusion would seem to follow. This move, however, may work against some will-centred libertarian views, but it begs the question against our proposal, which considers practical judgement as the central step in practical deliberation, and rationally controlled choice as derived from it. Assuming exact sameness between the

two agents until the very moment of choice, the difference between their respective decisions may be explained, in the context of our proposal, by pointing out that J*'s practical judgement, unlike J's, did not control his decision, which so was arbitrary and irrational. In fact, if J and J* form the same practical judgement, namely that not granting clemency, and so not raising their hands, is better than the opposite, and none the less J does not raise his hand while J* does, then J*, unlike J, acts against his best considered judgement and so in a weak-willed way. Being contrary to his best judgement, J*'s decision was irrational and arbitrary. But we cannot conclude that J's decision, just because it *might* also have been weak-willed, was equally irrational and arbitrary. Provided that J decided in accord with and owing to his practical judgement, his decision was rational and controlled by that judgement. That weakness of the will is always possible does not show that we never decide in a controlled way.

Even if an agent's practical judgement does not causally determine his decision, it is not mere luck that people's decisions usually accord with their practical judgements. When this is not the case, we need a causal/psychological explanation, which is not needed otherwise. The objection may implicitly assume that indeterminism undermines rational control by severing the causal link between practical judgement and decision. But the absence of causal determination does not erase the distinction between J's and J*'s decisions. A way of expressing the difference is to say that J's decision, but not J*'s, was justified, and so non-deterministically caused,[3] by the relevant practical judgement. In terms of our prior distinction, we might say that J*'s decision was a matter of bottom-up indeterminism, while J's decision was made within top-down indeterminism.

So, if the sameness between the two agents is assumed to hold until the very moment of choice, the conclusion that J's decision is arbitrary and chancy does not follow. The objector can respond to this reply by accepting that J and J* form different practical judgements. He may then try to reinstate the objection by pointing out that, under the assumption of indeterminism, J and J*, reasoning *in exactly the same way,* can form different practical judgements, which reveals the ultimate arbitrariness of J's actual judgement. Leaving aside Buridan cases, if, in the light of the same standards and evidence, and through exactly the same process of reasoning, including their mentally spoken words, J and J* form different practical judgements, this would seem to show that J's actual practical judgement was a matter of chance. But our proposal suggests a different perspective on the situation. Assuming *exact* sameness between the two agents until the forming of their respective practical judgements, if the latter differ, it seems that at least one of these judgements (presumably J*'s) was not appropriately backed by the relevant reasons. J*'s practical judgement was, perhaps, a case of weakness of warrant, and so an irrational phenomenon in itself, independently of the question of indeterminism. Again, that J *might* also have formed his practical judgement through weakness of warrant, as J* did, does not show that J's actual practical judgement, provided that it was rightly controlled by his reasons (the reasons he shared with J*'s), was affected by arbitrariness or irrationality.

The objector intends to conclude, from the fact that J and J* are exactly alike, and that, if indeterminism holds, they can form different judgements or make different decisions, that, under the assumption of indeterminism, our judgements and decisions are a matter of chance, so that indeterminism is incompatible with rational control. But the conclusion does not follow, as we have tried to show. There is at least one alternative explanation,

namely that J*'s practical judgement and/or decision, but not J's, were formed irrationally. It should be possible to distinguish, in an indeterministic world, between rational and irrational beliefs and decisions.

So our perspective can successfully meet the "Mind" objection if this is formulated in terms of exactly alike twins. In cases of less than perfect sameness, we can certainly accept that two almost alike agents can form different beliefs and decisions that can be equally rationally justified. The sameness that can be assumed to hold between the two subjects can go very far. It may extend to the holding of the same relevant evaluative beliefs, including of course the belief that capital punishment is justified, at least in some cases, as well as the same degree of competence in practical and theoretical reasoning. It may include many other psychological traits as well. With this degree of assumed sameness, we can accept that J and J* may well form different practical judgements and make different choices. But this does not force upon us the conclusion that the difference, and with it J's original decision, is just a matter of chance, even if indeterminism holds. For, if we are right that application of normative standards to particular cases leaves room for reasoned alternatives, J and J* may differ, for example, in the question whether the convict's crime falls within the scope of their common evaluative belief, that is, whether it actually is serious enough to deserve capital punishment, and so differ in their respective practical judgements *without this difference, and so J's decision, being a matter of chance*. Both judgements may be justified and under the agents' rational control, even if they cannot be both true.

Let us discuss the "Rolling-Back" version of the "Mind" argument, in Van Inwagen's presentation thereof. On the basis of the undetermined character of Alice's choice, and of the "Rolling-Back" thought experiment, Van Inwagen thinks we may plausibly conclude that, in the situation Alice faces, the actual result, namely her telling the truth, has the same probability, around 0.5, as the alternative result, namely her lying. So the fact that she actually tells the truth is a matter of chance and, we may add, not an adequate ground of her moral responsibility for that act. But why should we assign a probability of around 0.5 to each alternative? In making this move, Van Inwagen is taking a very external perspective of Alice's situation and thinking of indeterminism as a bottom feature of reality that transmits itself upwards, encompassing Alice's choice and action. But this view of both the situation and the role of indeterminism therein is not forced upon us. In connection with moral responsibility issues, we should rather make the effort to see things from the internal point of view of Alice herself when she faces the choice. In Van Inwagen's example, we are told next to nothing about Alice's point of view. We know almost nothing about her evaluative beliefs, especially concerning telling the truth and lying as kinds of actions, but this issue is central in our proposal. We may then imagine various possibilities.

Suppose, first, that Alice adopts a purely neutral stance, so that she does not think that telling the truth is either better or worse than lying. She does not think there is any difference in moral value between these two options either. Suppose also that, in this particular situation, considerations of another sort do not break the balance in favour of one of the options. This brings the situation close to a Buridan case. Here it is justified to hold the view that the probability of Alice's lying is more or less the same as that of her telling the truth, around 0.5. But this does not show that Alice lacks rational control over her choice for, in Buridan cases, it is a rational procedure to leave the decision to chance

by, say, flipping a coin. Any of the alternatives will then be justified. An additional question is whether she truly deserves praise or blame for what she chose and did, and the answer depends upon Alice's having or not having ultimate regulative control over her neutral (and plausibly false) evaluative belief.

However, Van Inwagen's remarks that the situation was "difficult" and that Alice "seriously considered telling the truth, [and] seriously considered lying" (Van Inwagen 2000:14) suggest that this reading of the example is not the one he has in mind. Alice is plausibly taken to assign moral relevance to her choice. Suppose that she has developed a system of evaluative beliefs that includes the general view that telling the truth, unlike lying, is morally good, though this value may be outweighed by other moral values, or even by non-moral ones, so that, in some situations in which this is the case, lying might be—even morally—justified. Suppose that her deliberation about the particular situation leads her to the view that, though there are some reasons for lying in this case, so that this option might have some degree of rational and even moral justification, these reasons do not seem sufficient to override the general evaluative belief that, *ceteris paribus,* telling the truth is preferable to lying. The practical judgement she arrives at, then, is that, in this particular situation, telling the truth is better than lying, and she decides and acts accordingly.

Are Alice's choice and action a matter of chance? What leads to this conclusion in the "Rolling-Back" argument is that, given that the choice is undetermined, in some subsequent replays Alice will choose to lie instead, so that after a high number of such replays the ratio of each possible outcome with respect to the total number of replays would be seen to converge towards a certain value, around 0.5. This means that, in the actual situation, there was a probability of about 0.5 that Alice had lied, so that her telling the truth is a matter of chance. However, after trying to see things from Alice's perspective, and to bring her evaluative views and practical judgement into the picture, various steps in the argument, as well as its conclusion, seem much less forceful. It is now important to ask at what phase in Alice's practical deliberation the replays are supposed to start in the thought experiment. Van Inwagen supposes they start "one minute before Alice told the truth" (Van Inwagen 2000:14). This is not much help, but it suggests that the replays are supposed to start at a time quite close to Alice's choice. Suppose they start at some moment between Alice's practical judgement and her decision. Then it is completely implausible to think that, in about 50 per cent of the replays, she will decide to lie and do so, unless we make the wild assumption that, in about half of the replays, she suffers an attack of weakness of the will. Suppose, then, that the replays start just before Alice's practical judgement. Then it is also implausible to think that, with the same reasoning that led her to her actual practical judgement in the real situation, she would arrive at a contrary practical judgement in about half of the replays. To get the argument off the ground, we should suppose that the replays start before she begins to reason about the alternatives, or perhaps in the middle of that reasoning. Assuming some degree of bottom-up, quantum indeterminism in Alice's brain, it may happen that in some of the replays some motives, related for instance to self-interest, appear to her as stronger than in the actual situation, thus leading her, in *some* of these cases, to a different process of reasoning and to a different practical judgement. It may also happen that in some replays she reasons more, or less, carefully, or takes other considerations into account, though this in itself does not mean that she will necessarily

arrive at a different practical judgement: our justified practical conclusions enjoy some degree of protection against such changes. So, even assuming that these variations take place in some of the replays, it is still implausible, given Alice's system of evaluative beliefs and her general mental traits, to think that the ratio of the two possible outcomes would take a value which suggested the conclusion that Alice's actual decision is a matter of chance. Alice's making the same choice as in the actual situation in subsequent replays is, we think, much more probable than her choosing otherwise.

What makes the sort of arguments we are considering seem unconvincing is the fact of facing them from the perspective of a reason-sensitive, top-down indeterminism, combined with a cognitive rather than volitive approach to moral responsibility. But indeterminism is still there. When Alice faces the choice between telling the truth and lying, what she will choose is undetermined. Her evaluative view and the data she has about the situation she faces do not determine her choice. She still has to reason about how the view applies to the particular situation, and about whether the latter falls within the scope of the former. It is in the nature of normative systems and practices that questions of this sort have open answers and leave room for different, and justified, judgements. Taking part in those systems and practices introduces into our life the reason-receptive indeterminism that is required for moral responsibility.

We may finally ask whether our evaluative beliefs, which, we have argued, make ultimate control over our choices and actions possible, are simply a matter of chance. It would be foolish to deny that the system of evaluative beliefs we start with largely depends on luck, an expression that is shorthand for factors beyond our possible control. But this does not mean that our praise- or blameworthiness for our evaluative views is equally a matter of luck. Blameworthiness for our evaluative views may diminish or even disappear if we do not possess the required control over the factors that explain such views. To take an earlier example, this is what happens in the case of Meno, the landowner in Ancient Greece, concerning his belief that slavery is morally permissible. Conversely, an agent's praiseworthiness for her correct views may be increased by virtue of the unfavourable circumstances in which she formed them. In any case, given minimally favourable circumstances, concerning both our innate capacities and our environment, we can control our evaluative views and get justification for them by applying the relevant normative standards. As happens with our free choices and actions, the evaluative system we hold at a certain time is undetermined, but again this indeterminacy does not threaten our rational control, for it relates to the reason-receptive indeterminacy which, we have argued, is constitutive of normative systems and practices.

Final remarks and Conclusion

We started this chapter with an attempt to diagnose the roots of scepticism about moral responsibility. A revision of the dialectic situation that results from the preceding four chapters showed that, while SMR's premise B (the incompatibility of determinism and moral responsibility) is supported by two main lines of argument, only one leads to SMR's premise C (the incompatibility of indeterminism and moral responsibility). This line leads to premise C through the necessity of ultimate control for moral responsibility and the incompatibility of indeterminism and ultimate control. Now, since ultimate

control seems clearly incompatible with determinism as well, then, if it is necessary for moral responsibility, the latter cannot exist. According to some authors, the incompatibility of ultimate control with both determinism and indeterminism is just a consequence of the fact that this condition is incoherent in the first place. This would seem to counsel a rejection of ultimate control as a condition of moral responsibility, as compatibilists have done. But we have taken a different route. Against compatibilists, we have defended the necessity of ultimate control for moral responsibility, understood as true desert; and, against the incoherence contention, we have tried to show that ultimate control is indeed possible. Now, since determinism seems clearly incompatible with that condition, given that there are no ultimate causal sources in a deterministic world, we have contended that ultimate control is none the less compatible with indeterminism. The main objection against this contention is that indeterminism turns our choices and actions into chancy and arbitrary occurrences, and so erodes our control (and a fortiori our ultimate control) over them.

We have tried to resist both lines of attack on ultimate control. We have argued for the hypothesis that what makes ultimate control falsely appear as an incoherent, and therefore impossible, demand is an approach to moral responsibility and its freedom-relevant conditions with an almost exclusive emphasis on choices and acts of will. And we have contended that it is also this approach that makes undetermined choices and actions falsely appear as matters of chance. On this basis, we have developed the makings of an alternative, libertarian approach, which instead underlines the central significance of cognitive factors and the existence of a form of control and true desert that does not give the will centre stage. We have especially insisted on the importance of evaluative beliefs for the possibility of moral responsibility. We have argued that, from this cognitive perspective, indeterminism can be seen to be receptive to rational control, against arguments to the contrary. This alternative approach, we have contended, clears the ground for an anti-sceptical stance towards moral responsibility, in that it undermines the central line of argument for SMR's premise C. However, humility is also an important virtue in philosophy, and it is especially important with regard to such recalcitrant questions as this book has been dealing with. More work should be done in order to fill in the details and reinforce the proposal against possible, and probable, objections. In the end, this proposal might well be shown not to work. In the meantime, however, let us enjoy the hope that it might succeed.

The cognitive approach to moral responsibility we have proposed in this chapter has important advantages over will-centred libertarian approaches, especially for what concerns rational control. However, we can also integrate important insights of both incompatibilist and compatibilist theories into this new context. We can accept, for example, with Kane, Benson and Richardson, that the freedom relevant to moral responsibility is not an inborn and indelible quality of human beings, but rather a contingent achievement that has to be acquired and perfected, and that can also be lost. We agree with Kane when he holds that moral responsibility concerns not only particular actions, but also the building of one's own character, of "virtuous and vicious dispositions through one's own efforts, choices, and actions" (Kane 1996:181), though we would also insist on the importance of cognitive virtues and attitudes, such as humility and respect for evaluative facts. And we can accept the relevance of external social and political factors to the development of one's freedom. We can allow for the

possibility of a plurality of justified systems of evaluative views, as well as for its importance in the formation of a correct system of our own. Moreover, we can avail ourselves of compatibilist insights about the factors that increase, diminish or even preclude moral responsibility, against a rather inhuman view of human beings as unrestrictedly free and responsible agents and so unconditionally liable to punishment. And, of course, we can gladly greet the compatibilist insistence on rational conditions of moral responsibility.

Conclusion

In the first four chapters of this book we have gone through the complex network of pathways that lead to the sceptical conclusion that moral responsibility is not possible. Our main task in those chapters has been to evaluate the sceptical argument about moral responsibility that we have dubbed "SMR". This argument's conclusion that moral responsibility is not possible is supported by three premises, namely: that determinism is either true or not true (premise A); that, if it is true, moral responsibility is not possible (premise B); and that, if it is not true, moral responsibility is not possible either (premise C). Since the argument is formally valid, our task has concentrated on the reasons for thinking that its premises are true. And, since premise A seems patently true, we have mainly focused on premises B and C.

One important line of support for premise B comes from the thesis that alternative possibilities are necessary for moral responsibility, together with the thesis that determinism rules out alternative possibilities. SMR's premise B follows from these two theses as premises. We have dubbed this argument "the Incompatibilist Argument".

In Chapter 1 we have dealt with the premise that determinism rules out, or is incompatible with, alternative possibilities of decision and action. Support for this premise comes from several sources. The most influential among them is the so-called "Consequence Argument", which has that premise as its conclusion. In the form that Van Inwagen gave it, this argument relies on a rule of inference, namely rule Beta, which has been shown not to be valid. But we have seen that there are still many other ways in which the incompatibility between determinism and alternative possibilities may be established. Rule Beta may be replaced by other rules that may well be valid. And such an incompatibility may be argued for in other forms, without resorting to rules of that kind. On this basis, we concluded, in Chapter 1, that the incompatibility in question is most likely true.

However, even if it is true that determinism rules out alternative possibilities, this in itself does not show that it rules out moral responsibility as well, unless alternative possibilities are required for moral responsibility. This is the other premise of the Incompatibilist Argument for SMR's premise B. Chapter 2 has been devoted to an assessment of this premise, which is also known as the Principle of Alternate Possibilities (PAP). According to PAP, an agent is morally responsible for an action of hers only if she could have done otherwise. This principle has been challenged in several ways, which we labelled as "Frankfurt cases", "self-trapping cases" and "Luther cases".

The first of these three lines has been by far the most influential, and many present thinkers are still convinced that it defeats PAP. We are not. Our starting point has been that PAP has a high degree of initial plausibility. As the principle that "ought" implies "can" (OIC), PAP is related to the control we understandably want to have over our moral responsibility for what we do. These two principles may well be logically connected: there are reasons to think that OIC, together with some plausible assumptions, implies PAP with respect to blameworthiness. If it does, then rejecting PAP would entail

rejecting OIC as well, and being led to accept very uncomfortable consequences. Corresponding to the initial plausibility of PAP, if Frankfurt cases are to be successful against this principle, they have to raise a strong intuitive judgement that the agent is morally responsible for what she does. We have gone through the main stages of the discussion—which is still going on—about Frankfurt-inspired attacks on PAP. "Flicker of freedom" (the term is Fischer's) theorists try to find alternative possibilities in Frankfurt cases that can ground the agent's moral responsibility. Fischer has found this strategy faulty in that those alternative possibilities, present as they may be, might not be "robust" enough to explain why the agent is morally responsible. He has distinguished two versions of this strategy: forward- and backward-looking, depending on whether the focus is on what would happen after the counterfactual factor's intervention or on the sign that triggers the intervention. Concerning the first version, since the agent does not act freely after the intervention, Fischer charges "flicker" theorists with "alchemy", an attempt to obtain freedom and moral responsibility from their absence. Concerning the second version, Fischer has designed cases where the triggering sign is a purely involuntary happening, so that the alternatives (for example, blushing or not blushing) are, in anybody's lights, not robust enough to account for the agent's moral responsibility. Fischer's move, however, has been challenged by a powerful defence of PAP, put forward by Widerker (and Kane). According to this defence, Frankfurt cases face a dilemma: either determinism is assumed to hold in the actual sequence of such cases or it is not; if it is, then incompatibilists will reject the agent's moral responsibility; if it is not, then it is hard to see why the agent's decision is unavoidable.

Nevertheless, the dilemma defence has not brought the discussion to an end. It has, however, forced strong restrictions on the construction of convincing Frankfurt cases. We have examined two main attempts to design cases immune to the dilemma: blockage cases and Pereboom-like cases. Blockage cases, of the sort proposed by Mele/Robb or Hunt, might be assuming determinism, so falling prey to one horn of the dilemma. But our main objection to them is that they violate a plausible rationality-related condition for moral responsibility, namely "reasons-responsiveness" (in Fischer and Ravizza's terms). Pereboom-like cases feature, as the triggering sign of the counterfactual intervention, a condition that is only necessary, and not sufficient, for an alternative choice. Cases of this sort seem able to avoid the dilemma, as well as the problems that affect blockage cases. However, we have argued that the "flicker" strategy can be successfully rescued and made to weigh heavily against cases of this sort. As a first step, we have tried to show that Fischer's "alchemy" objection to the strategy can be met. And we have further contended that the alternatives present in Pereboom-like cases are morally and explanatorily relevant for the agent's moral responsibility, and that, as a result of having them, agents can be shown to have more robust, "exempting" alternatives. We have generalized this contention to other Frankfurt cases. Though the discussion is still open, as we said, our provisional conclusion has been that PAP is safe against Frankfurt-inspired attacks on it.

A similar conclusion has been drawn concerning the two other lines of attack on PAP, namely "self-trapping" and "Luther" cases. Concerning the former, we have contended that, if subjected to a careful scrutiny, these cases can be shown to include alternative possibilities available to the agent. As for the latter, we have argued that the examples in which the alternatives are really unthinkable, which do not include Luther's case itself,

show that certain basic conditions must be fulfilled for being a morally responsible agent in the first place. They do not show that, once these conditions are met, an agent can be morally responsible while lacking any alternative possibilities.

The general result of Chapters 1 and 2 has been that there are strong reasons for thinking that the Incompatibilist Argument, whose conclusion is SMR's premise B, is actually sound.

In Chapter 3, we have examined a second argument whose conclusion is also SMR's premise B. The premises of that argument are that determinism rules out ultimate control and that ultimate control is necessary for moral responsibility.

The first premise seems clearly true, for, with the possible exception of a First Cause, there cannot be any ultimate causes or origins in a deterministic world: anything that we may take as cause or origin of something is itself caused or originated by something else as well.

It is the second premise, namely that moral responsibility requires ultimate control, which is really contentious, and Chapter 3 is largely devoted to a discussion of it. As we did with the alternative possibilities condition, we have argued that the ultimate control condition for moral responsibility is initially very plausible as well. This plausibility relates to the very nature and depth of moral responsibility ascriptions. These ascriptions are directed to the agent herself, on the assumption that she is the true origin of what she is held morally responsible for, and they deeply affect her personal worth and value. So, if the agent cannot be said to be the true, ultimate origin of what she is held morally responsible for, and to have a proper rational and volitional control over it, those ascriptions do not seem justified. The question is whether this second premise, in spite of its initial plausibility, could be shown to be false. We have defended this premise as well. Our defence has consisted in an examination of the main compatibilist attempts to justify moral responsibility without resorting to ultimate control. We have argued that, in spite of their valuable insights into many aspects of moral responsibility, none of these attempts is finally successful. This claim rests mainly on the fact that, against our deep intuitions to the contrary, all such approaches have to accept that agents in certain kinds of situations are morally responsible for their decisions and actions. Situations of these kinds include Watson's "Brave New World" cases and Kane's "CNC (covert non-constraining) manipulation" cases. Our intuition is that agents in these contexts are not morally responsible, and what seems to underlie this judgement is our perception that they lack ultimate control over their decisions and actions: they cannot be said to be their true origins and sources. It can be argued that, even in cases in which, unknown to the agent, her decisions and actions are induced by direct manipulation of her brain, these approaches should yield the result that she is morally responsible for such decisions and actions. And this is really very hard to swallow. Our conclusion, then, has been that ultimate control is actually necessary for moral responsibility, which, together with the premise that determinism excludes ultimate control, allows the conclusion that, if determinism is true, moral responsibility is not possible. And this conclusion is SMR's premise B.

Chapter 4 has examined SMR's premise C, namely that, if determinism is not true, moral responsibility is not possible, or, in other words, that indeterminism is incompatible with moral responsibility. Support for this premise comes from the contention just referred to, that ultimate control is a requirement for moral responsibility,

together with the assumption that indeterminism excludes control, and a fortiori ultimate control, over our decisions and actions. These two statements taken as premises imply SMR's premise C as a conclusion. Now, unlike determinism, indeterminism seems receptive to the possibility of ultimate causes or origins, in that it leaves room for events that are not determined by the past history of the world and the natural laws. These events, which might include our choices, could then act as fresh, ultimate starting points. The problem that indeterminism raises for the possibility of ultimate control, and so of moral responsibility, does not come, then, from the "ultimacy" aspect of that condition, but rather from its "control" aspect. That indeterminism rules out control over our decisions and actions can be seen as a corollary of the "Mind" argument's conclusion that undetermined events, particularly decisions and actions, are chancy, arbitrary occurrences. We have gone through several versions of this argument. Mele's "twins" version and Van Inwagen's "Rolling-Back" version are especially powerful.

We have taken Kane's theory of free will and moral responsibility as representative of libertarianism, in order to test it against the "Mind" objection. According to Mele, Kane's is "the most thoughtful and detailed defence of libertarianism currently available" (Mele 1999:96). I agree with this judgement, so our choice is not arbitrary. Central to Kane's theory are what he calls "self-forming actions" (SFAs), and especially "self-forming willings" (SFWs), undetermined choices among alternative courses of action backed by incommensurable sets of reasons, by means of which the agent forms her own self and motivational system. It is essential to the role that those SFWs should play that, in facing choices of this kind, the agent cannot be said to want to act on one set of reasons more than on the other, or to find one course of action more rational than the other, for otherwise those SFWs could not be said to be truly self- and will-forming and ultimate responsibility (Kane's construal of ultimate control) would not be possible. In order for this condition to be met, agents should have "plural" rational and voluntary control over their SFWs, so that, *whatever their choice is,* it should be shown to be rational and voluntary. We have tried to show that the way Kane proposes to meet this requirement makes his theory into a form of sheer voluntarism or decisionism, for in this theory it is *decision* that *makes* one set of *reasons* prevail over the other. At the end, only a completely baseless, and so arbitrary, choice could be truly ultimate and self-forming. The "ultimacy" and the "control" aspects of the ultimate control condition seem to pull in opposite directions.

On this basis, we have argued that, even if Kane's perspective can go some way towards meeting some version of the "Mind" argument, it ends up falling prey to others. Against the "twins" version, Kane resorts to indeterminate efforts of will and to quantum indeterminacy at the brain level in order to argue that the idea of the sameness of two individuals' psychological stories is not defined. This move is not unproblematic, but, even if it were actually successful against that version of the "Mind" argument, it does not seem able to meet the "Rolling-Back" version of it. Moreover, we have also tried to show that Kane's conception of efforts of will and choices in terms of indeterministic quantum-cum-chaotic events in the brain actually worsens the prospects of responding to the charge that indeterminism precludes rational control, for quantum phenomena are not in control of anyone, and if choices and will efforts are identical to, or supervene on, such phenomena, they will not be under anyone's control either. In the end, we have

concluded that, even if Kane's proposal can allow for ultimacy of source, it does not provide enough room for rational and volitional control.

So our examination of Kane's libertarianism has given strong support to the contention that indeterminism precludes control, and so ultimate control, over our decisions and actions. Moreover, as Van Inwagen has convincingly argued, other varieties of libertarianism, especially agent-causal approaches, do not seem to fare better against the "Rolling-Back" objection than event-causal ones, such as Kane's or Van Inwagen's. And if, as we have also accepted, ultimate control is necessary for moral responsibility, understood as true, objective praise- or blameworthiness, then the prospects for the truth of SMR's premise C, namely that, if determinism is not true, moral responsibility is not possible, are significantly improved. If premise B is true, as we have argued that it is, then, given that premise A is a logically necessary truth, we can conclude, and have actually done so at the end of Chapter 4, that the case for SMR's sceptical conclusion is very strong indeed.

In Chapter 5, however, we have tentatively explored the possibilities of resisting that sceptical conclusion. The necessity of ultimate control for moral responsibility plays a central role in supporting scepticism about moral responsibility. So a tempting and direct way of resisting scepticism is to reject the necessity of that condition. This would undermine SMR's premise C, and with it the sceptical conclusion of that argument. However, we have argued that following this path is not advisable, for it would not preserve moral responsibility understood as true, objective desert. Perhaps "moral responsibility" can still be given some other senses after rejecting the necessity of ultimate control, but not the crucial sense of true desert, of true praise- and blameworthiness, which is actually the target of sceptical attacks and the root of philosophical concern about such a putative property of persons.

Thus we have accepted the necessity of ultimate control. On this basis, our hypothesis has been that it is not ultimate control itself, but certain particular ways of construing and accounting for it, that are responsible for the appearance that this condition itself is incompatible with either determinism or indeterminism, and so impossible to satisfy. In other words, we have granted the intuition that some form of deep, ultimate form of control over our actions is required for moral responsibility, but have raised suspicions instead about certain theoretical accounts of such an intuition. A clue to what may be wrong with such accounts has been provided by Galen Strawson's sceptical argument for the conclusion that ultimate control (which he calls "true responsibility" or "true self-determination") is a logically impossible demand. The reason is that, in Strawson's own words, it requires "the actual completion of an infinite regress of choices of principles of choice" (Strawson 1986:29). If this is actually required by ultimate control, then it is not surprising that this condition is incompatible with determinism and also with indeterminism. An inconsistent, logically impossible demand is incompatible with any possible situation or state of affairs.

However, Strawson's argument rests on an implicit assumption, which can also be found in Kane's theory of free will and moral responsibility, namely that control is essentially a matter of *choice,* so that you cannot be said to control anything unless it is subject to your will and you can decide about it. In fact, this assumption has pervaded the discussion about freedom and moral responsibility. However, if we accept the ultimate control requirement for moral responsibility and we also assume that all control is

essentially grounded in acts of will, particularly choices, it is hard to see how we can escape the sceptical conclusion that moral responsibility is not possible. For, on such assumptions, either our choices are made on the basis of factors (Strawson's "principles of choice") that we have not chosen, and then we do not have ultimate control over such choices, or we are bound to choose those factors on no basis at all, and so our choice will be a matter of chance, as defenders of the "Mind" argument have always contended.

We have accepted that there is a constitutive link between moral responsibility and control, but we have disputed that there is also a constitutive link between control and choices or acts of will. Control may be based on choices, but it need not be. As a support for this contention, we have indicated examples in which we grant an agent control over a cognitive achievement of hers, and so praiseworthiness for it, even if such control does not rest on choices or acts of will. In fact, the achievement would lose value and its author would lose merit if she had made many aspects of it depend on her choice. Moreover, in these examples we can see that an agent may deserve unrestricted praiseworthiness for a cognitive achievement of hers even if many enabling conditions for the achievement were not of her own making. These examples concern an agent's beliefs, and it can be seen that they do not conform to Strawson's or Kane's conceptions of ultimate control. We have argued that control over one's beliefs does not involve choice, but rather an attitude of humility and respect to the facts. Choosing our beliefs is a way of losing control over them.

So, instead of construing the ultimate control requirement for moral responsibility in terms of choices, our proposal has been to construe it in terms of beliefs. Central among our "principles of choice" in matters of moral import are our evaluative beliefs. If evaluative beliefs are to make ultimate control over our choices and actions, and so moral responsibility for them, possible, they have to satisfy a number of conditions, to witness: they should be a central component of an agent's self; they should correctly be attributable to her as true author; they should potentially be under an agent's rational control and be causally effective on her practical judgements, choices and actions. We have argued that, under appropriate circumstances, evaluative beliefs may actually satisfy these conditions. So they seem initially able to play the required grounding role with regard to ultimate control and moral responsibility. We have also argued that the alternative possibilities, or regulative control, condition for moral responsibility applies to evaluative beliefs in as natural a way as to choices and actions: reflection on some examples seems to show that an agent is not morally responsible for certain of her evaluative beliefs (and the ensuing actions) if having relevantly different beliefs was completely beyond her reach. Nevertheless, control over our evaluative beliefs, we have insisted, is not essentially related to choice. As happens with our ordinary beliefs about matters of fact, it rather has to do with the right attitude of respect towards evaluative facts. It is related to our ability to *see,* rather than to choose or act.

A final and crucial condition that evaluative beliefs should meet is that rational control over them should be compatible with indeterminism. So we have tested our cognitive libertarian approach to moral responsibility against the traditional objection that indeterminism erodes rational control. As an important step towards meeting the objection, we have distinguished two ways in which indeterminism can affect our practical reasoning and decision-making. Whereas quantum indeterminism, in so far as it is amplified and transmitted to higher neurological and mental processes ("bottom-up"

indeterminism), can actually be said to erode rational control, the sort of indeterminism that is constitutive of normative, rule-governed systems and practices may actually be receptive to rational control and even essential for it. These systems and practices include our system of evaluative beliefs and our practical deliberations. Our evaluative beliefs are subject to normative standards such as coherence and responsiveness to experience; and evaluative beliefs, in turn, act as normative standards in our practical deliberations. At both levels, there is room for reasoned discussion about when a standard is actually met or what meeting it involves in particular cases. This is the sort of indeterminism that can sustain rather than erode rational control. We have tried to show how this reason-receptive indeterminism could act in a "top-down" direction, helping to shape the structure of "lower" levels of reality, such as the structure of the brain, without violating materialist intuitions about the supervenience of mental properties on physical ones.

On the basis of these considerations, we have finally come to face the most powerful versions of the "Mind" argument. Corresponding to the cognitive nature of our approach to ultimate (regulative) control and moral responsibility, practical judgements, instead of choices, are given centre stage in practical deliberation. Practical judgements, in turn, are aptly seen as applications of an agent's evaluative beliefs to particular situations she confronts. Choices, on this view, depend on practical judgements and remain linked to them. On these assumptions, we have argued that we are not obliged to accept the conclusion of the "twins" or the "Rolling-Back" versions of the "Mind" argument, according to which undetermined choices are a matter of chance. Concerning the "twins" version, we have contended that the possibility that an agent's actual choice or practical judgement may differ from those that an exactly alike twin of hers might make does not force on us the conclusion that the agent's actual choice and practical judgement are chancy, arbitrary events. In the context of our proposal, the discrepancy may be due to the fact that one of the twins chooses or forms her practical judgement in an irrational, weak-willed or weak-warranted, way. But this does not affect the other twin's rational control over her choice and judgement, if such there is. As for the "Rolling-Back" version, we have argued that, in terms of our proposal, there is no reason to think that, after a high number of replays of a particular decision-making process, the ratio between two possible outcomes (choices or actions) should take a value that would make the agent's actual choice and action appear as a matter of chance. If we take into account the agent's actual evaluative beliefs and practical judgements, the probability that, in subsequent possible replays, the agent chooses as she does in the original situation is plausibly seen as much higher than the probability of her choosing otherwise. These versions of the "Mind" objection may represent a strong threat against conative, will-or choice-based approaches to moral responsibility, but they seem much less disturbing for our cognitive proposal. Moreover, they assume a bottom-up conception of indeterminism; with a top-down conception in place, several steps in these arguments look much less forceful.

Considerations of this sort throw serious doubts upon the statement that indeterminism precludes control, and a fortiori ultimate control, over our decisions and actions. But, since this statement is an essential premise in the argument for SMR's premise C, we can also refuse to accept this premise, and with it SMR's sceptical conclusion that moral responsibility is not possible.

Let us finally add that our proposed approach to moral responsibility and its freedom-relevant conditions is opposed to some largely implicit and unexamined assumptions in many contemporary approaches to these matters. These assumptions are not casual, but instead reflect deep traits of our present Western culture. So, against a profoundly individualistic view of human beings, our approach instead stresses their social nature and their participation in normative systems as enabling conditions for their freedom and moral responsibility. Moreover, against a predominant view of human beings as essentially active interveners and decision-makers, reflected in those approaches' stress on choice and action, our proposal instead insists on the importance of a contemplative stance, on our seeing rather than on our deciding and acting in order to form correct and justified views of what is really worthwhile and valuable in our lives. The ability to form these views, which at certain places in this book we have called "wisdom", is crucial not only for our condition of free and morally responsible agents, but also for our happiness and the quality of our life.

Notes

Chapter 1

1 This, of course, is not the end of the story. Compatibilists may attempt to reject the incompatibility between determinism and alternative possibilities on the basis of other analyses of "could" or "ability". Incompatibilists are likely to object that these proposals are question-begging; but compatibilists will direct the same charge against those responses. A careful and sympathetic review of these compatibilist attempts, with original contributions, can be found in Kapitan (2002). Compatibilists have also argued against a central premise of the argument, namely the fixity of the laws, which would certainly look like an obvious truth. A first and important attempt was made by David Lewis (1981). More recently, Helen Beebee and Alfred Mele (2002) have argued that, on a Humean understanding of natural laws, there is a sense in which they might be said to be "up to us".

2 The discussion about the Consequence Argument is likely to continue in the near future. Van Inwagen (2004) has recently defended the incompatibilist conclusion of the Consequence Argument against some objections, especially those inspired by Lewis (1981).

Chapter 2

1 This sort of freedom is nicely characterized by Thomas Pink at the very beginning of his book *The Psychology of Freedom* and identified with freedom *tout court:* "By *freedom* I mean the freedom of alternative possibilities: the freedom to do things or not to do them, or—as I shall also put it—*control* over whether we do these things or not. It is just this freedom that we think we possess in relation to much of our action" (Pink 1996:1).

2 For a highly detailed defence of the thesis that OIC, together with some plausible assumptions, implies PAP (restricted to moral blameworthiness) see Copp 2003.

3 A related positive defence of PAP against Frankfurt-inspired attacks has been put forward by David Widerker (2003). Widerker's starting point is that for our moral disapproval of someone's behaviour the belief that she should not have done what she did is essential.

4 Another version of the dilemma defence of PAP can be found in Ginet 1996.

5 Widerker finds this view "puzzling given that he [Frankfurt] undertakes to establish a thesis such as IRR" (Widerker 1995:252, fn. 8).

6 Though the view that assuming determinism in the actual sequence of a Frankfurt-type example is question-begging against incompatibilists is widely shared among Frankfurt theorists, not all of them agree. Fischer (1999:112–14) contended that, on a certain construal of Frankfurt's argument for the conclusion that moral responsibility does not require alternative possibilities, the assumption might not be question-begging. More recently, Haji and McKenna (2004) have argued that not any incompatibilist is entitled to object that this assumption begs the question. Only those incompatibilists who think that determinism may

rule out moral responsibility for reasons other than its excluding alternative possibilities would raise the objection in a legitimate way. On the other hand, those incompatibilists who think that the only reason why determinism rules out moral responsibility is that it rules out alternative possibilities would not be entitled to raise the objection. Haji and McKenna acknowledge, however, that it would be reasonable for incompatibilists of this sort to demand that, in a Frankfurt-type example, alternative possibilities should be excluded only by virtue of the counterfactual factor. But this claim does not amount to the devastating charge of begging the question. I would tend to think that the assumption of determinism also begs the question in the latter case, for, as I have indicated, on that assumption, the alternative possibilities incompatibilist is asked to reject the conclusion of the Incompatibilist Argument as a preliminary step to her rejection of premise 1 of that argument. Things are rather complicated here, for it is not fully clear what begging the question amounts to in many cases. But I think that, in an argument that crucially depends on the intuitions raised by examples, assuming determinism is not good advice. Frankfurt, we may recall, insisted on keeping the question about causation of the agent's action separate from the question about alternative possibilities, so that what caused the action was not, at the same time, what deprived the agent of alternatives. In examples based on coercion, for example, the two questions get mixed up, with the consequence that, as in poorly designed scientific experiments, we cannot be sure which factor explains our judgement about the agent's responsibility. It would seem that introducing determinism in the actual sequence is also a way of mixing up the two questions. Compatibilists will insist that being deterministically caused to do something is different from being coerced into doing it. This has always been their central point. But, even if they concede the difference, incompatibilists will reply that the two things can be equated for what concerns their effects on moral responsibility. It is unlikely that this traditional stalemate will prove fruitful. So, question-begging or not, assuming determinism in the actual sequence of Frankfurt-type examples is hardly good advice for Frankfurt theorists. The question whether alternative possibilities are required for moral responsibility does not include the concept of determinism as such, and is best addressed independently of the latter.

7 If the twitch is not epiphenomenal, but, as a neurophysiological, contentless state, actually causes the agent's decision, this makes the case relevantly similar to the former, and even compatibilists may question the agent's moral responsibility.

8 In a recent paper, Derk Pereboom expresses a related worry about examples such as Fischer's, which feature signs so plainly irrelevant to the agent's moral responsibility as blushes or twitches: "…If the blush itself or something associated with the blush—perhaps Jones's having eaten a twinkie—deterministically explain his decision to kill, then anyone should be concerned that his action is being produced by something other than a normal deliberative process, which in turn raises the possibility that Jones is not morally responsible after all" (Pereboom 2003:192).

9 In previous versions (cf. Pereboom 2001:19), the necessary condition for Joe's failing to choose to evade taxes is a moral reason occurring to him with a certain force. Pereboom now takes this condition to be Joe's attaining a certain level of attentiveness to those moral reasons. With this change he intends to meet the objection that blameworthiness may require "that the agent understands that his action is morally wrong, which in Joe's case would seem to require some awareness of moral reasons" (Pereboom 2003:194).

10 There is, however, a worry that one might feel about the example in connection with the agent's responsiveness to reasons. As Pereboom depicts Joe's psychology, only a certain level of attentiveness to moral reasons could lead him to a decision not to evade taxes. This is a strange restriction, for what would happen if, for instance, Joe acquired the information that illegally claiming a tax deduction for local registration fees was going to be subject to a much more severe legal and financial action against claimants? This is a strong reason for not choosing to evade taxes, but it is not a moral reason. If Joe were not attentive to it at all,

this might mean that he is not adequately reasons-responsive, and his moral responsibility could be undermined. Now, the example might be amended so that a certain level of Joe's attentiveness to *any* reason for not evading taxes would trigger the device's activation, but this would seem to narrow down the agent's available space for free deliberation more severely than in the original example, perhaps too severely for the example to be convincing. For then the agent would have to reach his decision to evade taxes without paying serious attention to *any* reason against such a decision, and this may lead us to wonder whether the agent can be responsible for such an "automatic", untested decision. But maybe the example could be modified so as to dodge these difficulties. We shall assume, for the sake of argument, that it could.

11 It may be objected that, in fact, Helen had a third alternative, namely to lie. However, it would not be difficult to modify the example so as to exclude this alternative. Imagine that the device that administers TD is connected to a reliable lie detector, which responds to suitable changes in an agent's skin when she is about to tell a lie. So, if Helen tried to lie, TD would automatically be administered to her. I owe this objection to Stephen Laurence, in discussion.

12 We shall see later on that this description is, strictly speaking, mistaken: one cannot denounce someone else involuntarily. But we can leave this aside at this point in the discussion.

13 Actions of this sort—"necessarily intentional" actions, as we may call them—played an important role in the line of argument I developed in my 1990 book. I resorted to them in answering the "regress-problem", which threatened the very existence and even possibility of actions at all, as well as in providing a general criterion of agency (cf. Moya 1990:49–52). Necessarily intentional actions also play an important role in the context of providing a defence of PAP in the present book.

14 So, for instance, McKenna writes: "In the Frankfurt-type cases the alternatives are, either doing what one does of one's own intention, or being coerced into performing *the same kind of action* against one's will" (McKenna 1997:74; my emphasis).

Chapter 3

1 I developed this point further in Moya (1995). In that paper, I indicated a further related paradox that seems to follow from Frankfurt's view of control, namely that, on such a view, the more incoherent an agent's system of second-order preferences happens to be, the more likely it will be that any first-order desire on which she acts on a particular occasion is one that she reflectively endorses; and therefore the more likely it will be that, on any particular occasion, she acts freely or of her own free will, in Frankfurt's own terms. Thus it seems to follow from Frankfurt's view that an agent's freedom is directly proportional to the degree of incoherence that her system of second-order preferences actually has. And this looks like a paradox, or at least a very bizarre consequence.

Chapter 4

1 A rather different criticism of Kane's response to the "Mind" problem can be found in Almeida and Bernstein (2003). They insist on the presumed inability of Kane's theory to provide "contrastive explanations" of decisions or actions, and go on to contend that the theory cannot make room for the sort of control that moral responsibility requires. A contrastive explanation is one that not only explains why an event took place but also why that event took place *instead of another* which apparently was also possible.

2 A defence of libertarianism against the "Mind" objection has recently been put forward by Laura Ekstrom (2003). Ekstrom's approach does not rely on agent-causation, but remains, as Kane's does, within an event-causal perspective. Ekstrom's main negative target is Van Inwagen's (2000) line of argument for the conclusion that indeterministic events are chancy and not within the agent's control. The "Rolling-Back" argument is central to this line. Ekstrom contends that there is no single interpretation of the term "chance" on which the premises of Van Inwagen's argument are true. Moreover, she contends that libertarianism does not require contrastive explanations (see note 1, this chapter). Positively, she defends the compatibility between indeterminism and control. I think that the "Mind" argument survives Ekstrom's attack, but I cannot pursue the matter further.

Chapter 5

1 The assumption that all control and responsibility requires the will's intervention is very widespread. For an example, consider the following text by R.Jay Wallace: "Both negligence and recklessness can be taken to reflect *qualities of will,* as expressed in action, *and so* to be appropriate grounds for blame… Recklessness…involves a cavalier attitude towards risk that shows itself in the relation between one's choice and one's awareness of the risk in acting on that choice…and so recklessness can itself be *a blameworthy quality of the will.* Negligence and forgetfulness are slightly harder cases, perhaps, because there may not even be awareness of the risks involved at the time when one acts negligently or forgetfully. Here one may have to trace the moral fault *to an earlier episode of choice*… In this way, negligence and forgetfulness may also be traced to a *blameworthy quality of will*" (Wallace 1994:138–9, my emphasis). Examples of this assumption could be multiplied.

2 In a recent paper (Zhu 2004) I have found some interesting remarks, which point in the direction of my defence of a form of control not based on the will. The seemingly paradoxical notion of "passive action", which Zhu traces back to Frankfurt, involves the idea of control, though not voluntary. Non-interfering with a process can be a form of control over that process, whether or not it can rightly be taken to be an action. While driving, for example, there are several moments at which we do not *do* anything. None the less, it seems right to say that at those moments we are still driving and in control of our car.

3 I have argued elsewhere (Moya 1998) that justification implies (non-deterministic) causation. So, if a reason is to justify a decision or action, it has to cause it. On my view, in order to analyse reasons explanation correctly, we need not and should not conceive of the causation relation between reasons and action as independent of the internal coherence relationships

between the (content of the) reasons and the (description of the) action that are essentially involved in justification. Justification already includes a causal requirement. This view, I would think, is in a better position than Davidson's to avoid problems such as epiphenomenalism of mental properties and deviant causal chains.

References

Almeida, M. and Bernstein, M. (2003) "Lucky libertarianism", *Philosophical Studies,* 113:93–109.

Austin, J.L. (1970) *Philosophical Papers,* Oxford: Oxford University Press.

Ayer, A.J. (1954) "Freedom and necessity", in *Philosophical Essays,* London: Macmillan, 271–84; reprinted in (and quoted from) G.Watson (ed.) (1982) *Free Will,* Oxford: Oxford University Press.

Beebee, H. and Mele, A. (2002) "Humean compatibilism", *Mind,* 111:201–23.

Benson, P. (1994) "Free agency and self-worth", *Journal of Philosophy,* 91:650–68.

Burge, T. (1979) "Individualism and the mental", *Midwest Studies in Philosophy,* 4:73–121.

Chisholm, R. (1964) "Human freedom and the self", The Lindley Lecture, Department of Philosophy, University of Kansas; reprinted in (and quoted from) G.Watson (ed.) (1982) *Free Will,* Oxford: Oxford University Press.

Clarke, R. (1997) "On the possibility of rational free action", *Philosophical Studies,* 88:37–57.

Copp, D. (2003) "'Ought' implies 'can', blameworthiness, and the principle of alternate possibilities", in D.Widerker and M.McKenna (eds) *Moral Responsibility and Alternative Possibilities,* Aldershot: Ashgate.

Crisp, T.M. and Warfield, T.A. (2000) "The irrelevance of indeterministic counterexamples to Principle Beta", *Philosophy and Phenomenological Research,* 61:173–84.

Davidson, D. (1973) "Freedom to act", in T.Honderich (ed.) *Essays on Freedom of Action,* London: Routledge & Kegan Paul: 139–56; reprinted in D.Davidson (1982) *Essays on Actions and Events,* Oxford: Clarendon Press.

——(1982) *Essays on Actions and Events,* Oxford: Clarendon Press.

Dennett, D.C. (1984) *Elbow Room: the varieties of free will worth wanting,* Oxford: Clarendon Press.

Double, R. (1991) *The Non-reality of Free Will,* Oxford: Oxford University Press.

Ekstrom, L.W. (2000) *Free Will: a philosophical study,* Boulder (Colorado): Westview.

——(2002) "Libertarianism and Frankfurt-style cases", in R.Kane (ed.) *The Oxford Handbook of Free Will,* Oxford and New York: Oxford University Press.

——(2003) "Free will, chance, and mystery", *Philosophical Studies,* 113:153–80.

Finch, A. and Warfield, T.A. (1998) "The *Mind* argument and libertarianism", *Mind,* 107: 515–28.

Fischer, J.M. (1994) *The Metaphysics of Free Will,* Oxford: Blackwell.

——(1999) "Recent work on moral responsibility", *Ethics,* 110:93–139.

——(2003) "Responsibility and agent-causation", in D.Widerker and M.McKenna (eds) *Moral Responsibility and Alternative Possibilities,* Aldershot: Ashgate.

Fischer, J.M. and Ravizza, M. (1998) *Responsibility and Control: a theory of moral responsibility,* Cambridge: Cambridge University Press.

Frankfurt, H.G. (1969) "Alternate possibilities and moral responsibility", *Journal of Philosophy,* 66:829–39; reprinted in (and quoted from) *The Importance of What We Care About* (1988), Cambridge: Cambridge University Press.

——(1971) "Freedom of the will and the concept of a person", *Journal of Philosophy,* 68:5–20; reprinted in (and quoted from) *The Importance of What We Care About* (1988), Cambridge: Cambridge University Press.

——(1975) "Three concepts of free action", *Proceedings of the Aristotelian Society,* Suppl. Volume 49; reprinted in (and quoted from) *The Importance of What We Care About* (1988), Cambridge: Cambridge University Press.

——(1987) "Identification and wholeheartedness", in F.Schoeman (ed.) *Responsibility, Character, and the Emotions,* Cambridge: Cambridge University Press; reprinted in (and quoted from) *The Importance of What We Care About* (1988), Cambridge: Cambridge University Press.
——(1988) *The Importance of What We Care About,* Cambridge: Cambridge University Press.
Ginet, C. (1996) "In defense of the principle of alternative possibilities: why I don't find Frankfurt's argument convincing", *Philosophical Perspectives,* 10:403–17.
Haji, I. and Cuypers, S.E. (2001) "Libertarian free will and CNC manipulation", *Dialectica,* 55:221–38.
Haji, I. and McKenna, M. (2004) "Dialectical delicacies in the debate about freedom and alternative possibilities", *Journal of Philosophy,* 101:299–314.
Hookway, C. (1994) "Cognitive virtues and epistemic evaluations", *International Journal of Philosophical Studies,* 2:211–27.
——(2001) "Epistemic *akrasia* and epistemic virtue", in A.Fairweather and L.Zagzebski (eds) *Virtue Epistemology,* New York: Oxford University Press.
——(2003) "Affective states and epistemic immediacy", *Metaphilosophy,* 34:78–96.
Huemer, M. (2000) "Van Inwagen's Consequence Argument", *Philosophical Review,* 109:525–44.
Hume, D. (1975) *Enquiries Concerning Human Understanding and Concerning the Principles of Morals,* with introduction and analytical index by L.A.Selby-Bigge; revised text and notes by P.H.Nidditch; 3rd edn, Oxford: Clarendon Press.
Hunt, D.P. (2000) "Moral responsibility and unavoidable action", *Philosophical Studies,* 97: 195–227.
Kane, R. (1985) *Free Will and Values,* Albany, NY: State University of New York Press.
——(1996) *The Significance of Free Will,* Oxford and New York: Oxford University Press.
——(2000) "The dual regress of free will and the role of alternative possibilities", *Philosophical Perspectives,* 14:57–79.
——(ed.) (2002) *The Oxford Handbook of Free Will,* Oxford and New York: Oxford University Press.
Kapitan, T. (2002) "A master argument for incompatibilism?", in R.Kane (ed.) *The Oxford Handbook of Free Will,* Oxford and New York: Oxford University Press.
Lamb, J. (1993) "Evaluative compatibilism and the principle of alternate possibilities", *Journal of Philosophy,* 90:517–27.
Levy, K. (2001) "The main problem with USC libertarianism", *Philosophical Studies,* 105: 107–27.
Lewis, D. (1981) "Are we free to break the laws?", *Theoria,* 47:113–21.
McDowell, J. (1998) "Having the world in view: lecture one", *Journal of Philosophy,* 95:431–50.
McKay, T. and Johnson, D. (1996) "A reconsideration of an argument against compatibilism", *Philosophical Topics,* 24:113–22.
McKenna, M. (1997) "Alternative possibilities and the failure of the counterexample strategy", *Journal of Social Philosophy,* 28:71–85.
McKenna, M. and Widerker, D. (2003) "Introduction", in D.Widerker and M.McKenna (eds) *Moral Responsibility and Alternative Possibilities,* Aldershot: Ashgate.
Mele, A.R. (1999) "Kane, luck, and the significance of free will", *Philosophical Explorations,* 2:96–104.
Mele, A.R. and Robb, D. (1998) "Rescuing Frankfurt-style cases", *Philosophical Review,* 107: 97–112.
——(2003) "Bbs, magnets and seesaws: the metaphysics of Frankfurt-style cases", in D. Widerker and M.McKenna (eds) *Moral Responsibility and Alternative Possibilities,* Aldershot: Ashgate.
Moya, C.J. (1990) *The Philosophy of Action,* Cambridge: Polity Press.
——(1995) "A paradox in compatibilist accounts of free will and moral responsibility", *Crítica,* 27:119–27.
——(1998) "Reason and causation in Davidson's theory of action explanation", *Crítica,* 30: 29–43.
Naylor, M.B. (1984) "Frankfurt on the principle of alternate possibilities", *Philosophical Studies,* 46:249–58.

Nozick, R. (1981) *Philosophical Explanations,* Oxford: Clarendon Press.
Otsuka, M. (1998) "Incompatibilism and the avoidability of blame", *Ethics,* 108:685–701.
Owens, D. (2000) *Reason Without Freedom,* London: Routledge.
Parfit, D. (1997) "Reasons and motivation", *Proceedings of the Aristotelian Society,* Suppl. Volume 71:99–130.
Pereboom, D. (2000) "Alternative possibilities and causal histories", *Philosophical Perspectives,* 14:119–37.
——(2001) *Living Without Free Will,* Cambridge: Cambridge University Press.
——(2003) "Source incompatibilism and alternative possibilities", in D.Widerker and M. McKenna (eds) *Moral Responsibility and Alternative Possibilities,* Aldershot: Ashgate.
Pink, T. (1996) *The Psychology of Freedom,* Cambridge: Cambridge University Press.
Richardson, H.S. (2001) "Autonomy's many normative presuppositions", *American Philosophical Quarterly,* 38:287–303.
Schnall, I.M. (2001) "The principle of alternate possibilities and 'ought' implies 'can'", *Analysis,* 61:335–40.
Shah, N. (2002) "Clearing space for doxastic voluntarism", *The Monist,* 85:436–45.
Smilansky, S. (2000) *Free Will and Illusion,* Oxford: Clarendon Press.
Strawson, G. (1986) *Freedom and Belief,* Oxford: Clarendon Press.
Strawson, P.F. (1962) "Freedom and resentment", *Proceedings of the British Academy,* 48:1–25; reprinted in (and quoted from) G.Watson (ed.) (1982) *Free Will,* Oxford: Oxford University Press.
Taylor, C. (1982) "Responsibility for self", in G.Watson (ed.) *Free Will,* Oxford: Oxford University Press.
Unger, P. (2002) "Free Will and Scientiphicalism", *Philosophy and Phenomenological Research,* 65:1–25.
Van Inwagen, P. (1983) *An Essay on Free Will,* Oxford: Clarendon Press.
——(1994) "When the will is not free", *Philosophical Studies,* 75:95–113.
——(2000) "Free will remains a mystery", *Philosophical Perspectives,* 14:1–19.
——(2004) "Freedom to break the laws", *Midwest Studies in Philosophy,* 28:334–51.
Wallace, R.J. (1994) *Responsibility and the Moral Sentiments,* Cambridge MA: Harvard University Press.
Warfield, T.A. (2000) "Causal determinism and human freedom are incompatible: a new argument for incompatibilism", *Philosophical Perspectives,* 14:167–80.
Watson, G. (ed.) (1982) *Free Will,* Oxford: Oxford University Press.
——(1982) "Free Agency", in G.Watson (ed.) *Free Will,* Oxford: Oxford University Press.
——(1987) "Free action and free will", *Mind,* 96:145–72.
Widerker, D. (1987) "On an argument for incompatibilism", *Analysis,* 47:37–41.
——(1991) "Frankfurt on 'ought' implies 'can' and alternative possibilities", *Analysis,* 51: 222–4.
——(1995) "Libertarianism and Frankfurt's attack on the principle of alternative possibilities", *Philosophical Review,* 104:247–61.
——(2003) "Blameworthiness and Frankfurt's argument against the principle of alternative possibilities", in D.Widerker and M.McKenna (eds) *Moral Responsibility and Alternative Possibilities,* Aldershot: Ashgate.
Widerker, D. and McKenna, M. (eds) (2003) *Moral Responsibility and Alternative Possibilities,* Aldershot: Ashgate.
Williams, B. (1973) "Deciding to Believe", in *Problems of the Self,* Cambridge: Cambridge University Press.
Winch, P. (1963) *The Idea of a Social Science and its Relation to Philosophy,* 2nd edn, London: Routledge & Kegan Paul.
Wolf, S. (1990) *Freedom Within Reason,* Oxford: Oxford University Press.

Wyma, K.D. (1997) “Moral responsibility and leeway for action”, *American Philosophical Quarterly,* 34:57–70.
Zagzebski, L. (2000a) “Does libertarian freedom require alternative possibilities?”, *Philosophical Perspectives,* 14:231–48.
——(2000b) “Responses”, *Philosophy and Phenomenological Research,* 60:207–19.
Zhu, J. (2004) “Passive action and causalism”, *Philosophical Studies,* 119:295–314.

Index